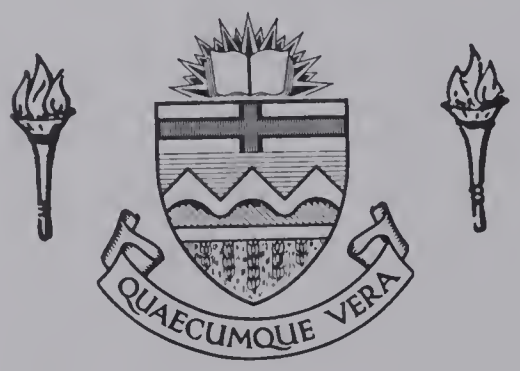


The University of Alberta
Printing Department
Edmonton, Alberta

UNIVERSITY OF ALBERTA

MEASUREMENT OF SIMILARITY OF VERBAL MATERIAL BY A PAIRED-COMPARISON PROCEDURE

by

WILLIAM RUSSELL MUIR

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

EDMONTON, ALBERTA

MARCH, 1968.

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a dissertation entitled "Measurement of Similarity of Verbal Material by a Paired-Comparison Procedure," submitted by William Russell Muir in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

A procedure was devised whereby an ordinal similarity scale of all possible pairs of stimuli in a set was derived for individual Ss from Ss' rating of a portion of the set of all possible pairs of stimulus pairs. This was done by supplementing the information from Ss' ratings with information provided by assumptions.

64 Ss rated for similarity two samples of verbal stimuli, each sample consisting of material of both high and low meaningfulness. Ss also practiced learning verbal responses to these stimuli for either 2, 4, 8, or 16 trials in a paired-associate learning (PAL) task. The similarity rating of the material was done both before and after the PAL procedure.

The reliability of the procedure, as assessed by a measure of internal consistency of response within a single rating session and by test-retest reliability of repeated ratings, was moderately high. Internal consistency of ratings was reliably affected by differences in Ss, specific samples of stimuli, and previous rating of stimuli at a different level of meaningfulness, but not by the meaningfulness of the stimuli or previous use of stimuli in a PAL task. Test-retest reliability was affected by the meaningfulness of the stimuli, but not by use of the stimuli in PAL between ratings. S showed highly stereotyped standards of similarity for low-

meaningful stimuli, but more idiosyncratic standards in rating high-meaningful stimuli. Rated similarity was reliably related to confusion errors after 4 trials, but not after 2, 8, or 16 trials of PAL, and to overt recall errors over the first 7 trials of PAL. Similarity rating of stimuli before PAL suppressed correct recall of responses to the stimuli during the early stage of practice.

Acknowledgements

I would like to thank the members of my advisory committee, Drs. W.N. Runquist, S.J. Rule, and W.W. Rozeboom for their advice and criticism on the planning, execution, and reporting of this experimental study. I would also like to acknowledge the invaluable aid of Messrs. R. Markley, K. Cranna, and S. Rankin in creating the computer program which made the analysis of the data of this study humanly possible.

My thanks are also due to the W.J. Gage Company Limited for their generous provision of a Research Fellowship to me during the academic year 1965-66. The freedom of being a Research Fellow during that period aided immeasurably the conception and execution of this study. I am also grateful for the aid received from the Principal's Research Fund of the University of Saskatchewan, Regina Campus, for expenses incurred during the writing of the report of this experiment.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

An important concept in contemporary verbal learning theory is that of similarity. Speculation as to the nature of similarity and its influence on human learning is at least as old as the British Associationists, and an attempt was made by Robinson as early as 1920 to use similarity as a construct in what would now be considered modern verbal learning theory and experimentation. Robinson's work was concerned with similarity of verbal material as a predictor of retroactive inhibition, still a topic of interest today. Later landmarks in the development of the concept of similarity in verbal learning were Gibson's (1940) use of stimulus generalization principles derived from conditioning experiments to explain phenomena such as intra-list interference in acquisition, Osgood's (1949) expansion and reformulation of the research sparked by Robinson to include transfer effects, and Bousfield's (1953) finding that subjects "cluster" similar responses in free recall. Current indications that interest in similarity has increased rather than waned are a recent book entitled Paired associates learning: The role of meaningfulness, similarity and familiarization (Goss and Nodine, 1965) and a chapter in a collection of readings surveying verbal learning (Kausler, 1966) devoted entirely to the topic of effect of similarity upon acquisition of verbal material.

With this increasing concern for similarity as a concept in verbal learning has come a growing awareness of a number of problems in attempting to operationally define and manipulate similarity. This thesis is

concerned with an examination of how similarity can be specified for verbal stimuli and with an experimental evaluation of a proposed procedure for scaling the similarity of verbal stimuli.

The Nature of Similarity in Psychology

During the last two decades a number of articles (e.g. Attneave, 1950; Noble, 1957; Wallach, 1958; Coombs, 1964) have examined the concept of similarity as used in psychology. This paper will not attempt an exhaustive review of these critiques, nor will it attempt to deal with various mathematical difficulties posed by attempts to measure similarity; rather, the discussion here will be restricted to a number of more or less general points which have been raised and which are particularly relevant to the measurement of similarity in verbal learning. The points discussed include the distinction between psychological and physical similarity, models of physical similarity, and the problem of multidimensionality in similarity relationships.

Psychological and Physical Similarity

Probably the most important distinction to be made in this regard is the one between psychological and physical similarity. When a learning theorist defines stimulus generalization as the tendency of an organism to give a conditioned response to stimuli similar to the training stimulus, or when a Gestalt psychologist states that similarity of stimuli induces perceptual grouping, it is almost invariably physical similarity that is referred to. Physical similarity is determined, as Wallach (1958) puts it, by the "Commonality of the environmental properties [which are] established by the identity of measurement readings of the centimeters-grams-seconds variety in two situations being compared" (p. 107). In

other words, physical similarity is defined by operations that are only incidental to a perceiving organism. For this reason Wallach refers to similarity defined in this way as "potential similarity", emphasizing that physical similarity is not necessarily related to psychological similarity.

Psychological similarity, on the other hand, can be loosely defined as the commonality of two situations as it is perceived by an organism, and thus it must be inferred from behavior. Although psychological similarity is not independent of physical similarity, it can be different from physical similarity for at least two reasons. (1) An organism may respond selectively to only a few aspects of a stimulus situation, and it need not respond equally to those aspects that it selects. (2) The responses that are elicited by stimuli in higher organisms are seldom fixed, as they are often dependent on previous learning. Thus, if we are to use the concept of similarity in any but the crudest of fashions in psychology, we must be prepared to look beyond the physical properties of the stimulus to the organism's mediating responses.

Models of Physical Similarity

Even attempting to specify the similarity of physical properties alone, however, presents problems. There would appear to be two main models on which the concept of physical resemblance can be patterned. The first is known as the common elements theory; it states that two situations resemble each other to the extent that they share identical constituent parts. The second theory, sometimes known as the dimensional model, assumes that one stimulus resembles another according to their proximity on some common attributive dimension. It can be seen that these models complement each other to a certain extent. The common

elements theory would intuitively seem most suited to the specification of the relationship between compound stimuli which can be broken down into easily-specified constituent parts (e.g. trigrams, which can be described as combinations of letter elements). The dimensional theory, on the other hand, would seem more suitable to the relating of stimuli possessing measurable properties such as objects of a particular size, hue, and brightness. Unfortunately, it is difficult to describe simple criteria, whether dimensional or elemental, for any but the simplest of stimuli. This brings us to the problem of determining similarity relations for stimuli that differ on a number of attributes.

Multidimensionality of Similarity Relations

James (1890) has pointed out that the moon is considered as similar to both a gas-jet and a football, but that the gas-jet in no way resembles the football. This is because the resemblance between moon and gas-jet is mediated by their common property of luminosity, while football and moon share the attribute of rotundity; however, James states, gas-jet and football have no properties in common. James' example is perhaps exaggerated -- gas-jets must have some recognizable shape, and a football must reflect a certain amount of light to be seen. Nevertheless, James emphasizes the importance of an aspect of similarity that has not always been recognized; that it is a multidimensional relationship.

Noble (1957) points out that similarity is not a dimension along which single stimuli can be scaled, a distinction that some researchers (e.g. Haagen, 1949) have not always made clear. Rather, it is an intransitive relationship between two objects; A may be similar in some degree to B, and B may be similar to C, but these statements tell us nothing about the similarity of A to C. This is because the similarity between A and B

can be mediated by those attributes or elements that A and B have in common, while the resemblance between B and C might be due to a set of attributes or elements that include all, some, or none of those common to A and B. This multidimensional aspect of similarity must be considered in any attempt to manipulate similarity experimentally.

In summary, previous discussions of the general nature of the concept of similarity in psychological theory have pointed out a number of areas where clear definitions of terms and relationships, although difficult, are necessary. These are the operational definition of physical similarity, the connection between physical and psychological similarity, and the multidimensional nature of the similarity relationship.

Measures of Similarity Used Previously in Verbal Learning

A wide variety of models of similarity, both physical and psychological, have been used previously in verbal learning studies. The procedures used to scale similarity have also varied widely, ranging from direct rating techniques where subjects are instructed to provide judgments of similarity through to what Flavell (1961a) terms "functional equivalence" measures. In the latter the similarity of the two stimuli is defined as the degree to which one stimulus can apparently replace another in some stimulus-response relationship. Examples of functional equivalence measures typically involve phenomena such as generalization, transfer of training, clustering in free recall, and interference. Although the functional equivalence measures have less intuitive appeal as indicants of similarity than other procedures such as direct rating, they are important not only because they have occasionally been used directly

as a defining operation for stimulus similarity (e.g., Gibson, 1940), but also because functional equivalence measures have frequently been used as a criterion by which to evaluate other types of similarity measure.

This section reviews, to the writer's knowledge, all the major methods used previously to scale similarity among verbal stimuli. The studies cited, however, are intended to illustrate the scaling procedures rather than exhaustively review all previous research. The measures described include three variations of a common elements model (with letters, phonemes and associations as elements), direct scaling of stimulus pairs, latency of similarity judgments, measures involving number of similarity criterion responses, estimates of the co-occurrence of the referents of meaningful words, and the semantic differential scaling technique, all of which involve direct ratings of stimuli. Also described are a number of measures based on functional equivalence of stimuli.

Common Elements

Three scaling methods are based on the assumption that stimulus similarity is mediated by common elements. Two of these are measures of the physical similarity of stimuli, considering letters and phonemes as elements, while the third measures psychological similarity by considering as elements association responses to stimuli.

Letter elements. Underwood has used the technique (e.g., Feldman and Underwood, 1957) of constructing lists of trigram stimuli from a limited number of letter elements, with similarity due to element duplication increasing with decreasing numbers of constituent letters. Abbott and Price (1964) used a 3-letter nonsense syllable as a training

stimulus in an eyeblink conditioning procedure, then tested subjects' responses to the same syllable and to nonsense syllables with two, one, or no letters in common with the training stimulus. In a more sensitive form of this procedure, Runquist and Joinson (in press) varied not only number of common elements between pairs of nonsense syllables but also the position of the repeated letters in the sequence.

Phoneme elements. Conrad (1962, 1964) and Wickelgren (1965, 1966) have studied the effects of acoustic similarity of letters and digits as measured by common phonemes in experiments where material was presented visually or aurally. Strictly speaking, this is a measure of physical similarity for aurally-presented stimuli only; the phonemic similarity of visually-presented verbal stimuli is mediated by a learned vocal response. However, for purposes of this analysis it would seem to do no harm to consider measures of phonemic similarity as being essentially equivalent for visually and aurally-presented stimuli.

Associative response elements. The third common-elements method is a measure of psychological, rather than physical, similarity. It treats as constituent elements of a stimulus the association responses elicited by it in an association test. The basic procedure is to calculate the degree to which association responses to two stimuli overlap; e.g., stimuli A and B might elicit few or no association responses in common, indicating that their similarity is low, while stimuli C and D might elicit many of the same responses (including each other), indicating high similarity. It is beyond the scope of this paper to describe the large number of different procedures that exist for calculating an index of similarity by this method (Marshall and Cofer, 1963; Goss and Nodine, 1965; Flavell and Johnson, 1961); however, the principle

of overlap or commonality of associative responses is essentially the same for all of them.

Direct Scaling

In a number of studies subjects have directly rated the similarity of pairs of stimuli by various scaling methods. McGeoch and McDonald (1931) had judges place 60 pairs of synonyms into three equal groups on the basis of decreasing similarity, and then picked the ten most consistently judged pairs from each group to represent three degrees of intra-pair similarity. Haagen's (1949) subjects used a seven-point scale with similarity defined as "the extent to which words denote the same or similar objects, actions, or conditions" (p. 454). The stimuli rated were common two-syllable adjectives, previously arranged in sets of six as being similar in dictionary definition. One member of each set was selected as a focus, and median ratings of the similarity of this focal word to the other five members were calculated for 80 subjects. In a study that rated trigrams of various types and association values rather than meaningful words, Runquist and Joinson (in press) had subjects give direct magnitude estimates of numbers between 0 and 100 to represent the similarity of pairs of stimuli. Garskof and Houston (1963) used a similar procedure for scaling pairs of meaningful words except that their subjects responded by marking an "X" on a 5-inch line with ends marked 0 and 1.

Latency

Flavell (Flavell, 1961b; Flavell and Johnson, 1961) presented subjects with pairs of stimuli and asked them to think of some way in which the stimuli were similar. The time period between presentation of the pair and the subject's report of some basis for perceiving the

stimuli as similar was used as an inverse measure of similarity.

Production of Similarity Criteria

Flavell and Johnson (1961) had subjects write as many similarities between a given pair of stimuli as they could in one minute. Subjects were told that the similarities need not be familiar or logical, but should be "genuine ways in which the referents were alike" (p. 339). A number of indices were derived from these data and also from the data of the single dominant responses reported in the latency procedure described above. All of these indices estimated similarity either through number of criteria produced per stimulus-pair or through measures of stereotypy of criteria between subjects.

Co-occurrence of Referents

Flavell (Flavell, 1961b; Flavell and Johnson, 1961) had subjects give probability ratings of the co-occurrence of the referents of meaningful words (adjectives, concrete nouns, and abstract nouns) in the environment. These probability estimates were used as measures of the semantic similarity of the pairs.

Semantic Differential

The semantic differential technique (Osgood, 1952; Osgood, Suci, and Tannenbaum, 1957) was originally constructed as a method for measuring the meaning of single stimuli. Based on Osgood's (1953) mediating response theory of meaning, it assumes that the meaning of a verbal stimulus can be specified by its position on a number of scales defined by bipolar adjectives, e.g., "fast-slow", "good-bad", "strong-weak". Factor-analytic studies have shown that most of the variance of the individual adjective-scales could be accounted for by three main semantic factors. These factors have been used by Osgood and his associates to describe a

three-dimensional semantic space, within which the meaning of verbal symbols could be specified by their projections on the three sets of defining coordinates. Assuming that this semantic space is Euclidean, similarity of meaning of two stimuli can be equated with the distance between the two points representing the two stimuli as determined by the generalized Pythagorean theorem (Osgood and Suci, 1952).

Functional Equivalence

This approach attempts to specify the similarity of two stimuli by measuring the extent to which one stimulus can replace the other in a behavioral relationship. In other words, similarity is scaled, not directly, but by observing some aspect of behavior that has a well-established theoretical relationship with similarity. Examples of phenomena through which this method has been tested are stimulus generalization in classical and instrumental conditioning, transfer of training, intra-list interference in acquisition and clustering in free recall.

Stimulus generalization in classical conditioning. Razran (1949) classically conditioned a salivation response to a meaningful word, and then measured the amount of conditioned response to test words of varying degrees of semantic and phonemic similarity. Riess (1940) measured generalization of a conditioned galvanic skin reflex to meaningful words, while Abbott and Price (1964) conditioned an eyeblink response to a trigram, then tested subjects' responses to the training stimulus and other trigrams.

Stimulus generalization in instrumental conditioning. Dicken (1961) trained an instrumental lever-pulling response to a set of words, then tested subjects with another group of words and noted the frequency

of response to each. Postman (1951) presented subjects with a number of 6-letter nonsense bi-syllables, then had them attempt to recognize the previously-experienced stimuli from among a larger group of bi-syllables. Similarity of a given class of bi-syllable was determined by its relative frequency of being "recognized".

Transfer of training. A typical study in which similarity was measured by transfer was done by Bastian (1961). Subjects learned first one paired-associate list, then learned a second list containing identical stimuli and different responses, followed by additional practice on the first list. Similarity of the response members was indicated by the amount of positive transfer associated with each response. A similar design was used by Ryan (1960), except that the similarity of the stimulus members rather than the response members was measured.

Intra-list interference to paired-associates learning. Wimer (1963) estimated the average similarity of groups of meaningful words through intra-list interference by using them as stimuli in paired-associates lists and determining the mean number of trials that subjects required to learn each list. Feldman and Underwood (1957), in addition to measuring average similarity of stimulus and response members in paired-associates lists by trials to criterion, also used mean overt errors per trial as an additional measure of similarity.

Clustering in free recall of lists. Bousfield, Whitmarsh and Berkowitz (1960) measured the frequency of co-occurrence of pairs of words in free recall after the list of words had been presented for learning in a randomized list. "Clustering" was used as a measure of similarity.

In conclusion, a review of measures of similarity relations among

verbal stimuli has been presented. It was shown that under one classification system the measures could be classed according to the assumptions underlying the experimental operation by which similarity was specified. These assumptions can be described dichotomously: those measures based on physical similarity and those based on psychological similarity; and dimensional versus common element measurement procedures. An alternative classification system groups measures of similarity into three classes according to the type of measuring operation; those based on intrinsic properties of the stimulus (the letter and phoneme common element measures), those based on subjects' direct responses to the stimuli (associative overlap, direct scaling, latency, production of similarity criteria, and the semantic differential), and measures based on subjects' responses to stimuli when they are part of a more complex task (the functional equivalence measures).

The Locus of Similarity

From the above review it would seem that the majority of similarity measures have implicit in them the assumption that similarity relations between stimuli can be studied directly through the properties of the verbal stimuli themselves. Some measures presumed that the similarity of stimuli could be specified in terms of their physical characteristics. Other measures were based on phychological similarity; i.e., as mediated by the properties of responses elicited by the stimuli. However, most measures of psychological similarity seem to assume that the response to a given stimulus is sufficiently stable and stereotyped within cultural groups that common standards may be applied to individuals within the group (e.g., Osgood, 1953). Despite the fact that Osgood (1962) reports finding inter-individual differ-

ences in semantic differential ratings of words within cultural groups, most studies have apparently tacitly assumed that differences in semantic structure between individuals can safely be ignored in studies of similarity. Garskof and Houston (1963) showed that a measure of similarity based on individual subjects' data gives a more reliable relationship with stimulus generalization than does a measure based on pooled data from many subjects. However, their study seems to be a lone exception to the rule that individual differences have received little attention in the study of similarity relations between verbal stimuli.

It is suggested that the shift in emphasis from the study of similarity through the invariant properties of stimuli to the study of the responses of the individual organism to the stimuli would greatly clarify the problem of individual differences. This approach was suggested by Osgood in 1953, but does not seem to have been seriously or consistently considered in verbal learning studies, even by Osgood himself. Briefly, it is presumed that an environmental stimulus can produce a mediating response in the organism's central nervous system. The nature of this response is determined by a number of factors--the sensory processes involved, the organism's previous experience with the stimulus, and the aspects of the stimulus being attended to, to mention only a few. These mediating responses have stimulus properties to which overt responses can be attached, resulting in such phenomena as stimulus generalization and transfer of learning as a function of the similarity of the mediating responses.

This mediating response hypothesis redefines the dichotomy between primary and mediated or secondary generalization; it states that all generalization is mediated in that it is a function of the properties of responses in the central nervous system to the stimuli. According to this view what

is called primary generalization occurs when the mediating response is determined directly by the sensory processes elicited by the stimuli, while secondary generalization occurs when the mediating response to the sensory process is modified by learning. Thus, psychological similarity can be studied at three stages. First, we may study the events antecedent to the mediating responses, including the physical properties of the stimuli, the organism's prior experience with the stimuli, and how it attends to the stimuli. Second, the mediating responses themselves may be examined; and third, the consequent behavior involving these mediating responses can be measured.

It can be seen that some of the measures of similarity reviewed above can be arranged on a continuum, the extremes of which represent the antecedent and consequent links of these mediating responses. Various measures differ in that they attempt to measure similarity at different points on the continuum. The phoneme and letter common-element measures tap the similarity process at the antecedent end by assessing the physical characteristics of single stimuli and then postulating the nature of the psychological similarity relationship. In moving toward the consequent end we next encounter the associative overlap and semantic differential measures. Both these measures first study what might be considered as the properties of mediating responses to single stimuli and then specify hypothetical similarity relationships. At the consequent end of the dimension we have the functional equivalence measures which assess the degree to which one stimulus can substitute for another in stimulus-response relationships.

The remaining scaling techniques (direct scaling, latency measures, similarity criteria production, and referent co-occurrence) do not fit as

obviously into this continuum as do the others. These measures have one aspect in common that differentiates them from other similarity measures; namely, that similarity is scaled by the strength of a response elicited by a pair of stimuli presented to a subject. The other measures present single stimuli to the subject and relate the independently elicited responses by some indirect procedure, so that the measure of similarity is derived from the data. In the four techniques listed above, however, the similarity of a pair of stimuli is directly inferred from the subject's response to that pair. One of these measures, co-occurrence of referents, assumes that similarity of verbal symbols is caused (at least partly) by previous association of the referents of these symbols in the subject's environment and is an attempt to scale the degree of this association. Because of this feature co-occurrence of referents is probably best considered as a variety of associative overlap measure. However, the other three measures entail no assumptions as to the causes of similarity relations among verbal stimuli. These measures (direct scaling, latency, and criteria production) are concerned with similarity as an empirical, rather than as a theoretical, function. They assume only that the psychological similarity of two stimuli can be scaled directly by subjects in much the same way as attitudes or preferences, and thus are best considered as operational definitions of similarity.

This leads to the central point of this thesis. Do these operational measures of similarity, especially direct scaling, which are essentially attempts to psychophysically scale subjects' subjective estimates of similarity, offer any advantage to the psychologist that is not found in the antecedent and consequent measures of similarity? In terms of experimental convenience and of relating the theoretical construct to the in-

tuitive concept of similarity, the answer would appear to be yes. It is obvious that a simple scaling procedure for determining similarity relations would represent a saving of effort and parsimony relative to procedures such as the antecedent and consequent measures of similarity described earlier. And the direct measures of similarity have the added attractiveness that they are direct manifestations of human observers' standards of similarity. In discussing what should be the primary data for similarity relations, Coombs has stated

> It has always seemed self-evident to me that the observations should be verbal judgments of similarity. We could, of course, utilize a transfer experiment and observe changes in amplitude or latency and use such observations to measure similarity. Even if this were done, however, the ultimate criterion for accepting the measure as a measure of similarity would be subjective. It seems to me that verbal judgments should be used to construct a measure of similarity, and then psychological theory would revolve around the functions that would relate similarity to other behavior phenomena. (Coombs, 1964, pp. 436-437).

Granted that the empirical measures of similarity have some intuitive appeal, it must now be shown that they are also useful. That is, it must be shown that direct ratings of similarity are reliably related to other events; specifically, to the antecedent conditions that presumably cause similarity and to the consequent behavior that is affected by similarity. A number of experiments indicate that empirical measures of similarity are related to antecedent and consequent measures of similarity. Representative studies of these relationships will be reviewed in the following sections, together with examples of studies relating different empirical measures of similarity.

Relations between Antecedent and Direct Measures of Similarity

In this section some representative studies relating antecedent and direct measures of similarity are summarized. They have been grouped

according to whether the antecedent measures were derived from physical, semantic differential, or associative commonality techniques of scaling similarity.

Physical measures of similarity. Runquist and Joinson (in press) had subjects rate pairs of nonsense syllables for similarity. They found that similarity ratings could be predicted from the number and position of shared letter elements in the pairs. Using two different sets of nonsense syllables their data showed the mean similarity ratings yielded a correlation of rho = .977 when pairs constructed according to the same principles were compared.

Semantic differential measures of similarity. Flavell (1961b) found correlation coefficients between judged similarity and similarity as determined by the semantic differential ranging from r = .86 for pairs of adjectives to r = .40 for adjective-concrete noun pairs. Wimer (1963) found that the correlation between judged similarity of nouns and semantic differential similarity (summation across 20 scales) was r = .547.

Associative commonality measures of similarity. Wimer (1963) obtained correlations between judged similarity of noun-pairs and four measures of associative overlap. Two of these correlations were significant, those with "total associative overlap" (r = .405) and "associative reciprocity" (r = .434), while the correlations with "associative overlap within individuals" and "variety of associative overlap" had moderate but nonsignificant correlations with ratings of similarity. Haagen (1949) found a correlation coefficient of r = .90 between ratings of pairs of adjectives for "closeness of associative connection" and "similarity of meaning". The former measure differs from most scales of associative commonality in that subjects rated pairs of stimuli instead of producing associations to single stimuli. It

is possible that Haagen's high correlation between associative commonality and rated similarity is due to subjects' operating on rated similarity of meaning as an indicant of associative connection. However, Cofer (1957) obtained a more orthodox associative overlap index ("mutual frequency") for pairs of Haagen's stimuli. Although Cofer did not report a correlation between Haagen's (1949) similarity ratings and his own associative commonality scores, it would appear from his published results (Cofer, 1957, p. 605) that a correlation of rho = .915 between rated similarity and associative overlap can be calculated from his grouped data.

An important experiment was that of Garskof and Houston (1963) in which two measures of associative overlap were correlated with each other and with rated similarity of pairs of nouns. One measure of associative commonality was "mutual frequency", used previously by Cofer (1957). This measure involves only the first association response given by a subject to a stimulus, and requires group data for the calculation of an index. The second measure of associative commonality used by Garskof and Houston was the "relatedness coefficient", which involves a weighting of the sequence of association responses given by a subject to a stimulus. This second measure can be used to calculate indices of associative overlap for individual subjects. Garskof and Houston found correlations between the "relatedness coefficient" and rated similarity of 24 word pairs ranging from rho = .63 to rho = .94 among 20 subjects, all significant at the .01 level of significance. The rank order correlation over all 20 subjects between the "relatedness coefficient" and rated similarity was rho = .94. In contrast, "mutual frequency" for 13 of the 24 word pairs as calculated from group data was 0, and the correlation between "mutual frequency" and rated similarity for the 11 pairs which did obtain a "mutual frequency"

greater than 0 was negative and nonsignificant. Correlation between "relatedness coefficient" and rated similarity for the same 11 pairs was rho = .97. This is apparently the only published study in which measures of similarity based on data from individual subjects have been calculated. Garskof and Houston suggest that the discrepancy between their nonsignificant results and Cofer's (1957) significant findings concerning the relation between "mutual frequency" and rated similarity may be due to the different numbers of subjects employed (20 and 356, respectively). Nevertheless, this experiment would seem to support the contention that similarity relations may be measured more accurately through single subjects than through considering pooled data from a group of subjects.

Relations between Direct and Consequent Measures of Similarity

In this section will be reviewed representative studies relating direct measures of similarity to consequent measures of similarity. These studies will be grouped according to the types of behavioral phenomena that were used as consequent measures of similarity; namely, intralist effects in paired-associates learning, interlist transfer, generalization of semantic conditioning, and confusions in short-term memory.

Intralist effects in paired-associates learning. A number of studies have shown that rated similarity between stimuli or responses can affect various aspects of the acquisition of a paired-associates list. Some representative studies involving stimulus similarity, response similarity, and a "concept-learning" design with similar stimuli and repeated responses will be described.

In a typical experiment studying stimulus similarity Underwood (1953) had subjects learned paired-associates lists whose stimulus members varied in similarity according to Haagen's (1949) norms. He found a complex

relationship between stimulus similarity and rate of learning, with medium-similarity lists taking longest to learn. Wimer (1963) constructed six-pair lists composed of six low-association responses used in all lists and 32 common-noun stimuli, each stimulus used in six lists. She recorded the mean number of trials to master each list to one errorless trial and the mean similarity rating for all pairs of stimuli on each list. The correlation between trials to criterion and mean list similarity was $\underline{r}$ = .428.

Among experiments varying rated similarity of response members in paired-associates lists was one by Higa (1963). He found that a list containing as responses pairs of words rated as synonyms was more difficult to learn than a control list. Underwood (1953) found that lists containing similar responses produced rates of overt errors during acquisition that were proportional to the degree of response similarity.

Richardson (1958) had subjects learn a number of 16-pair lists consisting of two groups of eight stimulus members with high intragroup rated similarity in all lists. The lists varied as to the number of different response members that were used and also as to whether repeated responses were paired with similar or dissimilar stimuli. In this way learning a list with repeated responses to similar stimuli was analogous to learning a concept. It was found that pairing a repeated response to similar stimuli produced positive intralist transfer.

Interlist transfer. A number of studies have shown that rated similarity of response members between lists causes positive transfer of training from first to second list. Bastian (1961) found significant transfer effects in second list learning when pairs of response items in the two lists were judged to be semantically similar by a direct scaling technique. Both Underwood (1951) and Morgan and Underwood (1950) found that interlist

response similarity defined by Haagen's (1949) norms caused facilitation of second-list learning and instrusions of first-list responses during learning of the second list. Slamecka (1967) showed that rated synonymity of responses in two lists causes positive transfer in both mixed and unmixed list designs.

Generalization of semantic conditioning. Several studies have shown that when a response has been trained to a verbal stimulus, generalization of responding is shown to other stimuli judged similar to the training stimulus. This has been shown for both classical and instrumental conditioning procedures.

Typical of the classical conditioning studies was Razran's (1939) in which a salivation response was conditioned to four words. Subjects were then tested for response to synonyms of the training stimuli. He found significant generalization of responding. Riess (1940) replicated Razran's study using the galvanic skin response instead of salivation, with comparable results.

Kurcz (1964) instrumentally conditioned a key-pressing response to word stimuli. She found that subjects showed generalization to synonomous words.

Confusions in short-term memory. Baddeley (1966) aurally presented sequences of words to subjects who then wrote down the sequences from memory after short delays. He found that using sequences of words that were judged similar to each other produced a small but reliable decrement in recall of the correct sequence.

Relations between Different Empirical Measures of Similarity

A few studies have shown that different empirical measures of similarity are related. Flavell and Johnson (1961) found judged similarity

of concrete nouns correlated r = .57 with production of similarity criteria and r = .70 with latency, indicating significant relationships. Attneave (1951) found a correlation of r = .91 between judged similarity of concrete nouns and criteria production. There is some evidence, then, that judged similarity is related to the other direct measures of similarity. It should be noted that there is a large difference between Attneave's (1951) and Flavell and Johnson's (1961) reported correlation between judged similarity and production of similarity criteria. Flavell and Johnson point out that their experiment used two independent groups for the two measures of similarity whereas Attneave had the same subjects undergo both procedures. This supports the hypothesis that similarity relations may be idiosyncratic to a given subject.

Summary and Conclusion

It has been suggested that similarity is better considered as a relationship between internal representations of external events than as a direct relation between the external events themselves. Although schemes have been proposed by which environmental events or objects are considered as similar or dissimilar according to their intrinsic properties, it has been shown that usually some criterion must be arbitrarily picked to determine which properties of stimuli will be used to determine their similarity. In addition stimuli are often judged as similar, not on their own intrinsic properties, but according to properties of responses that are elicited in the observer.

If similarity is considered in this way, then an illuminating distinction can be drawn between the different types of similarity measures used in verbal learning. Three classes of similarity measure have been proposed in this paper. The first type studies the properties of environ-

mental events and the operations that cause two stimuli to be perceived as similar. These typically involve a theory of the mechanism involved in similarity relations. The identical elements measures imply that stimuli are similar because they share similar sensory components. These components are either the letters that make up the visual representation of the verbal stimulus or the sounds of the vocal responses the stimuli habitually elicit. These measures assume that the similarity of the stimuli is determined by their sensory properties. Although the semantic differential and the associative overlap measures assume that similarity relations are a function of the covert responses (meaning) the organism has learned to the stimuli, they have been classed with the physical measures here because they specify theoretical mechanisms that determine the similarity of stimuli. In the case of the semantic differential the meaning of the verbal stimulus is the representational mediation process it elicits which is part of the total behavior originally provoked by the object for which the verbal stimulus is a sign. The associative overlap measures assume that a given verbal stimulus elicits covert verbal responses in addition to the one that is isomorphic to the pronunciation of the given stimulus.

Another type of similarity measure--the consequent measures--assumes that stimuli perceived as similar will affect responses learned to these stimuli in a certain way. Any phenomenon that is theoretically related to stimulus similarity can be used to construct a measure of this class--stimulus and response transfer, stimulus differentiation in first-list learning, and clustering of responses in free recall of a list are only a few of the measures that have been or could be used.

The third type of similarity measure--direct rating--could be considered as a variety of consequent similarity measure. However, direct

ratings have certain distinctive features that would seem to warrant putting them in a category of their own. Whereas the consequent similarity measures all are based on the subject's behavior toward the stimuli in a learning or perceptual task, the direct rating measures involve a psychophysical judgment by the subject that depends on his use of the predicate "is similar to." This type of similarity measure was referred to as an empirical measure, in that it is an operational definition of similarity entailing no assumptions or description of the nature of similarity. It was shown, however, that there is evidence indicating these direct measures of similarity, which can be regarded as subjects' psychophysical assessments of the relations between their mediating responses to stimuli, are reliably related with many of the antecedent and consequent measures of similarity.

It would seem that these empirical measures of similarity and their relations with other types of similarity measure are a potentially fruitful field for research that deserves detailed study. Two lines of argument support this contention. The first is that it would seem worthwhile to determine if subjects' impressions of similarity relations are reliably related to the antecedent and consequent conditions of similarity. The evidence reviewed earlier suggests this is the case; however, a valuable contribution could be made by providing a more precise technique for scaling perception of similarity.

The second argument concerns the present lack of cohesiveness in the attacks on the antecedent and consequent aspects of similarity. Each of the theories specifying how antecedent events affect the perception of similarity might be valid; similarity might be caused by a variety of causes, or there might be more than one variety of similarity. As presently formulated, however, the antecedent measures of similarity are limited in

the range of material that they are relevant to. It would be difficult for an associative commonality measure to be used over a wide range of stimulus meaningfulness, since by definition low-meaningful stimuli usually elicit few associative responses. Although acoustic similarity has been shown to have reliable effects on the learning and retention of high-meaningful material (Dallett, 1966), experiments comparing physical and semantic similarity seem to show that the dominant feature determining the similarity of high-meaningful stimuli is their meaning (Razran, 1939; Abbott, 1966), which limits the usefulness of the sensory component measures. And research (Osgood, 1962) seems to indicate that the semantic differential only assesses one limited aspect of meaning. Concerning the consequent measures of similarity, these involve complex phenomena the measurement of which pose difficult problems in themselves. It is entirely possible, for instance, that the psychological distances between a set of stimuli will be perceived differently by a subject when he is attempting to discriminate between the stimuli for the first time while learning a list than when he is attempting to recall the same stimuli after learning the list to some criterion.

It is the thesis of this paper that subjects' direct ratings can provide measures of similarity that relate to antecedent and consequent conditions of similarity. It is also believed that scaling techniques are best suited to obtain estimates of similarity relations unbiased by other effects, and that in addition they provide the most convenient means of obtaining data on similarity relations. If this presumption is not immediately evident from previous data on similarity rating, it might be explained in two ways.

First, previous similarity measures have almost invariably assumed some basis for similarity. These assumptions were often explicit, as in

the semantic differential and associative overlap measures. But upon examination of the instructions used in many of the direct rating studies it is obvious that an implicit assumption has been involved. Phrases such as "similarity of meaning" or "referring to the same things or actions" typically are used to clarify to the subject the nature of the experimental task. These instructions would also probably sensitize or set the subject toward certain attributes of the stimuli and produce ratings which might be limited in usefulness.

The second point is that all previous studies of direct similarity ratings have used the pooled data from several subjects. Often this was done of necessity as the reliability of the measures used was too poor for use on individual subjects. What has not apparently always been realized is that this procedure is based on the assumption that the commonality of similarity structures between individuals is sufficiently great that ". . . different subjects may be regarded merely as independent and random replications of each other . . . " (Coombs, 1964, p. 435). Coombs points out that this is not the case if, for instance, the stimuli are color patches and some subjects are color blind. It is interesting to note that Helm and Tucker (1962) found individuals with normal color vision differed in the three-dimensional psychological structures derived from their color-rating data. When it is considered that meaning, one of the most important attributes of verbal stimuli, is dependent on learned responses, then it would seem reasonable to attempt to verify the assumption that meaning structure is stable from individual to individual. The large difference found by Garskof and Houston (1963) between the correlations of measures of associative overlap based on group and individual data with similarity ratings and the difference between the correlation between rated similarity

and similarity criteria production found by Attneave (1951) with his intra-group procedure and Flavell and Johnson's (1961) intergroup procedure also suggests that better results might be possible using data from individual subjects.

In conclusion, this study will explore the hypothesis that individual variations in similarity structures exist for verbal stimuli, and that it is possible to assess individual similarity structures and the theoretical relations between similarity and performance in verbal learning. A proposed method for accomplishing this aim will be outlined in the following section.

Individual Measurement of Similarity Relations

The main obstacle to attempting to obtain direct ratings of the similarity of verbal stimuli for individual subjects appears to have been the amount of labor required of the subject and the experimenter to get reliable data by previous scaling methods. Osgood et al (1957), for example, in commenting on a study by Rowan (1954) comparing measures of similarity obtained by the semantic differential and the method of "triads" (presenting the subject with three stimuli and asking him to indicate which of the three possible pairs of stimuli show the greatest or least degree of a given relationship) say "The method of triads and other methods of the same type are excessively laborious and time-consuming . . ." (p. 145). Although they acknowledge that the method of triads involves fewer assumptions and a more spontaneous basis of judgment, Osgood et al conclude that "the 'freer' but more laborious method of triads should be used to test the validity of more 'restrictive' but simpler methods like the semantic differential" (p. 146), rather than as a practical scaling method for experimental work.

If only scaling operations involving many repetitions of the same similarity ratings in order to obtain stable estimates of stimulus distances (e.g., Torgerson, 1952) are considered, then Osgood's estimate of the method of triads as impractical is probably correct. However, Coombs (1964) has suggested a method of analyzing data from the method of triads whereby a single estimate of each triad can produce an ordinal scale of interstimulus distances. Coombs points out that the assumption of transitivity of similarity relations provides information which enables the experimenter to demand less of the subject in the rating task. For example, if the similarity of stimuli A and B is rated as greater than the similarity of stimuli B and C, and if B and C are judged more similar than C and D, then by the assumption of transitivity A and B should be more similar than C and D. Coombs discusses scaling methods which provide enough redundancy to test the assumption of transitivity, and yet enable the experimenter to arrange the experimental task so as to obtain a large amount of information about stimulus similarity from a moderate amount of effort on the part of the subject.

Only two studies seem to have utilized Coombs' suggestions for the scaling of similarity of verbal stimuli. Vastenhouw (1962) had subjects rate for similarity 17 names of personality traits. He found that subjects were consistent in their ratings and that information from a limited number of ratings could be used to predict the rated similarity of combinations not previously judged. He also reported that subjects' predictions of the probability of co-occurrence of personality traits in persons was related to their similarity ratings of the traits. Dember's (1957) subjects scaled the words "always", "often", "rarely", "seldom", and "never". The response latency of judgment was proportional to the semantic distance between

stimuli. However, the stimuli had been chosen so as to be described by but a single dimension of meaning. Thus, although Coombs' procedure seems promising for the scaling of verbal stimuli, no extensive test appears to have been made of the technique for this purpose.

Effects of Familiarization

Presumably, testing an idiographic approach to similarity scaling would necessitate a within-subjects experimental design. It can be seen that the scaling procedure and the validation operation must involve the same subjects if it is assumed that each subject has his own idiosyncratic structure of similarity relationships. This poses the problem of whether being exposed to the stimuli in the scaling procedure before the validation test would affect performance on the validation operation, or whether the converse would happen if similarity scaling were done after the validation procedure. If a learning task is involved in the validation procedure the process of familiarization would be involved.

Familiarization refers to a group of experimental procedures in which a subject receives a controlled amount of exposure to material prior to attempting a learning task involving this material. It has been hypothesized (Gannon and Noble, 1961) that amount of previous experience with a word determines its meaningfulness; hence experimental familiarization of material before including it in a learning task should increase its meaningfulness and the ease of learning the material through acquired distinctiveness. There has been controversy over the nature of the theoretical mechanism that supposedly mediates this acquired distinctiveness. In their review of work done on discrimination learning, Tighe and Tighe (1966) conclude that there are two main schools of thought concerning the cause of

acquisition of distinctiveness of cues. The differentiation hypothesis has it that with practice the organism becomes more sensitive to distinguishing cues within the stimuli. The mediation hypothesis assumes that the organism supplements the originally similar cues in the stimuli with distinguishing covert responses, thus rendering the stimuli more distinctive.

However, a further aspect of the problem complicates the matter even more. Operations which have frequently been used to specify familiarization procedures--repetition or inspection of the experimental material--have also been used to produce the phenomenon of semantic satiation, or decrease in meaning of verbal stimuli (e.g., Lambert and Jakobovits, 1960). To confuse the issue even further conflicting data have been reported both from experiments designed to investigate familiarization and from experiments designed to investigate semantic satiation--both types of study have produced increases and decreases in meaningfulness. Amster (1964) has reviewed the evidence concerning these conflicting effects in satiation studies and concludes that many of the data can be accounted for by variation in experimental procedures and materials. Goss and Nodine's (1965) review of the effects of stimulus familiarization on subsequent learning suggests that the equally contradictory results found in this field may have resulted from the welter of familiarization procedures as well. It would appear that closer control of the stimulus exposure operations will be necessary to determine how the mechanisms of familiarization and satiation work.

It can be seen that a stimulus rating procedure can be considered as a stimulus exposure operation, and that having a subject rate a group of stimuli before attempting a learning task could possibly affect performance on this task. Conversely, it is conceivable that learning responses to stimuli in a paired-associates list might affect the rediating responses

to the stimuli and hence the ratings of similarity relations between the stimuli (cf. Staats and Staats, 1957). Alternatively, it might be expected that learning to associate distinctive responses to a group of stimuli could cause the subject to attend more to the distinguishing aspects of the stimuli, and cause a change in similarity ratings. In any case it would be of interest to examine subjects' performance on associating responses to stimuli that had previously been rated for similarity, and also to test to find if using a group of stimuli in a paired-associates learning task changes the perceived similarity relations among the stimuli.

Statement of the Problem

No method has previously been tested for estimating the rated similarity of verbal stimuli for an individual subject. As the properties of verbal material determining their perceived similarity would seem to be caused by responses elicited by them in the organism, and as the nature of the response to a given stimulus would depend on the organism's previous experience, it follows that an idiographic measurement technique would be necessary to determine the similarity relations between stimuli which different individuals might have experienced in different ways. This study will attempt to validate a technique for measuring the similarity of verbal material over a wide range of meaningfulness for individual subjects. This will be accomplished by determining both the reliability of the proposed procedures and the relationship between rated similarity of stimuli and stimulus confusion errors in a paired-associates learning task.

A secondary purpose of this investigation will be to study the effects of stimulus familiarization (produced by a similarity-rating operation) on acquisition in subsequent paired-associates learning. A related

secondary purpose will be to study the effects of learning distinguishing responses to stimuli upon the perceived similarity of those stimuli.

CHAPTER II

Method

The general plan of the experiment was that subjects first rated an experimental and a control set of stimuli for similarity, then attempted to learn responses in a paired-associate task to half of these stimuli plus an equal number of new stimuli for a set number of trials. On the last trial of the learning task a recognition test was substituted for the recall test. Finally, the subjects repeated the similarity ratings for all the stimuli that had been used in the paired-associate test, but not rated previous to the learning task.

Material

Two paired associate lists (Table 1) were constructed, each list containing 12 pairs. The same 12 two-syllable adjectives were used as responses in both lists. The response terms were selected so that no two began with the same letter, no words had any obvious semantic relation to each other, and all had an AA rating in Thorndike and Lorge's (1944) list (signifying that they were among the 1,000 most frequently occurring words in written English and so are presumably highly meaningful in the commonly used sense of the word). These measures were taken to minimize the effect of response similarity during the learning of the lists. The stimulus items were selected according to two main principles:

1. Each list contained two equal groups of homogeneous stimuli representing two distinct and non-overlapping levels of meaningfulness.

2. Within each meaningfulness level there was a broad range of inter-stimulus similarity. The low-meaningful stimuli were consonant-consonant-consonant (CCC) nonsense syllables from Witmer's (1935) list which corresponded to no known English words or abbreviations. All syllables had Witmer association scale values of 17% or less. The mean association value for the CCCs in List 1 was 10.0%, and for those in List 2, 12.2%.

TABLE I

STIMULUS AND RESPONSE

TERMS OF PAIRED-ASSOCIATE LISTS

Stimuli		Responses
List 1	List 2	Lists 1 & 2
DJZ	HOLY	EARLY
WZQ	NICE	ABOVE
CGP	LAZY	BROKEN
ZMJ	DARK	HEAVY
HZJ	RICH	WEARY
ZXJ	CALM	FAMOUS
HARD	CXK	READY
FULL	XGK	SIMPLE
MILD	ZKG	OFTEN
SOFT	WGP	PUBLIC
SLOW	MHB	TENDER
WISE	GQK	COMMON

The high-meaningful stimuli were four-letter adjectives taken from Thorndike and Lorge's (1944) list. Two (both from List 1) were estimated to

be among the approximately 500 most frequently occurring words in written English, five (two from List 1, three from List 2) as among the first 1,000, three (one from List 1, two from List 2) from the first 2,000, one (List 1) from the first approximately 2,800, and one (List 2) from approximately the first 3,400 most commonly occurring words.

From the above specifications it would seem reasonable to assume that the stimulus material represents two discrete groups of items on the dimension of meaningfulness, while the homogeneity of meaningfulness within the groups is quite high. The reasonableness of the assumption of a large inter-group difference would seem assured by the fact that half the items are virtually unpronounceable combinations of letters that have been selected for their low probability of eliciting any associations from subjects, while the other half are extremely common English words. Technically, the assumption of high intra-group homogeneity is not as obviously reasonable, for the meaningfulness index of the CCCs ranges from 0 - 17%, while the frequency of occurrence of the highly meaningful material represents extremes of from once per 500 words to once per 3,400 words. However, an examination of the extremities among these stimuli (ZMJ and WGF versus ZXJ in the CCCs and FULL and HARD versus LAZY in the meaningful words) lends little intuitive support to the argument that appreciable intra-group differences might exist in the functional meaningfulness of these stimuli for highly literate subjects, relative to the inter-group difference.

Choosing the stimuli so as to ensure a wide range of inter-stimulus similarities on dimensions other than meaningfulness was done in two different ways. As it was assumed that the dominant feature determining inter-stimulus similarity for the CCCs was formal similarity, the nonsense syllable stimuli in the two lists were selected to meet the following criteria:

(1) The initial and final leters in two stimuli were the same.

(2) A third stimulus contained the same three letters as one of the stimuli in (1), but with the first two letters transposed.

(3) A fourth stimulus contained the two letters common to the first three stimuli, but in different positions.

(4) A fifth stimulus had the same middle letter as the stimulus in (3).

(5) The sixth stimulus had no letters in common with other stimuli.

(6) All letters apart from those specified in (1) to (5) were different.

In summary, in six stimuli one letter was repeated five times, one letter was repeated four times, one letter was repeated twice, and seven letters were used only once. The repeated letters occupied a variety of positions and combinations of positions in the stimuli.

It was presumed that the dominant feature to which subjects react in meaningful inter-stimulus similarity is their semantic properties. The meaningful stimuli were selected from those words in the Semantic Atlas published by Jenkins (1960) describing the scores on the three principal factors of the semantic space found by Osgood et al, (1957) that also appeared in Thorndike and Lorge's (1944) list of words with high frequency of occurrence in written English. Assuming that these three factors divide the semantic space into eight octants, six stimuli were chosen for each list that met the following criteria:

(1) Two stimuli were from the same octant.

(2) Three stimuli were from three other octants.

(3) One stimulus was from near the point of origin of the three principal factors, i.e., was relatively "neutral" by the criteria of the Semantic Differential.

It is hoped that this procedure would produce a wide range of inter-stimulus ranges, and that the two lists would be as comparable as possible as to inter-stimulus differences. However, for the high-meaningful stimuli it is realized that the second objective could be reached only in the most approximate of fashions, for a variety of reasons.

(1) Although the criteria for selecting octants were the same for both lists, a different pattern of octants were sampled for the two lists. Only one stimulus from List 2 came from the octant that yielded two stimuli for List 1; there was no stimulus on List 1 from the octant where the two most similar stimuli on List 2 originated.

(2) Little faith was put in the reliability of the inter-stimulus differences determined from the Semantic Atlas for the reasons discussed earlier in this thesis.

(3) Although one stimulus on each list was supposedly "neutral" in that its three principal axis scores were all close to the point of origin in the semantic space, it is probably more likely that these stimuli, as Jenkins (1960) points out, draw inconsistent and therefore cancelling responses from subjects. This conclusion is supported by the fact that although the two "neutral" words involved, FULL and DARK, are approximately contiguous in the semantic space with the concept GOJEY (a meaningless "paralog"), they are also quite close to CONSCIENTOUS OBJECTOR and SOCIALISM, concepts that would presumably be reasonably meaningful for the American college students who were the standardization sample for the Semantic Atlas. It is presumed, then, that FULL and DARK will not be neutral or central points, but will show extremely large inter-subject variations as points in semantic space.

It might be asked why the Semantic Atlas was used to select these stimuli, if it is known that its reliability is so poor. The problem here,

however, is not to get an accurate rating of the inter-stimulus distances but to ensure that the sample of stimuli selected represents an adequate range to test the proposed procedure. Although the Semantic Differential would seem to be an imperfect instrument for measuring inter-stimulus distances, it appears to be the best means available of assuring a wide range of inter-stimulus distances among highly meaningful verbal stimuli. It must be realized, of course, that less is known about the differences between the high-meaningful stimuli than about the distances between the low-meaningful stimuli, and so there is less assurance that the high-meaningful stimuli on List 1 will be roughly comparable to the high-meaningful List 2 stimuli.

The 12 response terms were paired at random with the 12 stimulus items of List 1 and the 12 stimulus items of List 2, the only restriction being that the responses which were paired with the high-meaningful material on List 1 were paired with the low-meaningful material on List 2, and vice versa. Three presentation orders of the 12 stimulus-response pairings and three recall test orders of the 12 stimuli alone were arranged for the two lists. The orders were randomized, with the restriction that the stimuli in the last two stimulus-response pairs in a given presentation could not occur in the first two stimuli in the immediately ensuing recall test. The same arrangements of stimulus-response pairings and stimuli were used in Lists 1 and 2, equivalence between items on the two lists being decided by the common response.

The recognition test for any given subject consisted of 24 stimulus-response pairings, with each stimulus and response in the paired-associate list occurring twice in the test. A number of forms of the recognition test were used. To construct the different tests (Table II),

TABLE II

PAIRINGS OF STIMULI AND RESPONSES USED ON RECOGNITION TEST

Stimuli								Responses
TEST A				TEST B				
List 1		List 1		List 1		List 2		
First Half	Second Half	First Half	Second Half	First Half	Second Half	First Half	Second Half	
DJZ[a]	MZJ	HOLY[a]	RICH	CGP	DJZ[a]	LAZY	HOLY[a]	EARLY
ZXJ	DJZ	CALM	HOLY	WZQ[a]	ZMJ[d]	NICE[a]	DARK[e]	ABOVE
MZJ[b]	CGP[a]	RICH[c]	LAZY[a]	ZMJ	WZQ	DARK	NICE	BROKEN
ZMJ[a]	WZQ[d]	DARK[a]	NICE[e]	ZXJ	MZJ[f]	CALM	RICH[g]	HEAVY
WZQ	ZMJ[f]	NICE	DARK[g]	MZJ[a]	CGP[b]	RICH[a]	LAZY[c]	WEARY
CGP	ZXJ[a]	LAZY	CALM[a]	DJZ	ZXJ[a]	HOLY	CALM[a]	FAMOUS
HARD[a]	SLOW	GXK[a]	MHB	MILD	HARD[d]	ZKG	GXK[a]	READY
WISE	HARD	GOK	GXK	FULL[a]	SOFT[i]	XGK[a]	WGP[e]	SIMPLE
SLOW[h]	MILD[a]	MHB[k]	ZKG[a]	SOFT	FULL	WGP	XGK	OFTEN
SOFT[a]	FULL[i]	WGP[a]	XGK[e]	WISE	SLOW[j]	GQK	MHB[m]	PUBLIC
FULL	SOFT[j]	XGK	WGP[m]	SLOW[a]	MILD[h]	MHB[a]	ZKG[k]	TENDER
MILD	WISE[a]	ZKG	GQK[a]	HARD	WISE[a]	GXK	GQK[a]	COMMON

[a] correct pairings

[b-m] pairs of "mirror image" mismatchings

four six-by-six matrices of all the possible combinations of stimuli and responses in the four groups of six stimulus-response pairs were drawn up, each comprising six correct and 30 incorrect possible stimulus-response pairings. From each of these matrices two basic recognition tests were constructed, each test being divided into two halves. (Two forms of the test were used to ensure that while a broad sample of mispairings was tested, the individual subject was not overburdened. The test was divided into halves to separate the two occurrences of each stimulus and response as widely as possible, and to allow for counterbalancing to control sequence effects.)

Latin squares were used to impose the following conditions upon these tests:

(1) Each half-test contained two correctly paired and four mismatched stimuli and responses.

(2) Each stimulus and each response was used once and only once in each half-test.

(3) Stimuli and responses that had occurred as correct pairings in a given half-test were used in mismatched pairs in the complementary half-test.

(4) No particular stimulus-response mismatching was repeated on both halves of a test.

(5) The two forms of the recognition test were made as independent as possible of each other in the following way:

(a) Only two of the six possible correct stimulus-response pairings were repeated. These two repetitions were unavoidable, of course, as there were eight correct pairings used in all.

(b) No stimulus-response mismatching was repeated on both tests.

(c) From the correct stimulus-response pairs $S_i - R_i$ and $S_j - R_j$ two mismatched pairs, $S_i - R_j$ and $S_j - R_i$, can be constructed. These two mispairings might be considered as "mirror images" of each other. Only three of the mismatched pairs presented to a given subject were mirror images of other mispairings. Thus, of the 16 stimulus-response mispairings involved in a given six-by-six matrix, 13 may be regarded as orthogonal to each other, and three to represent mirror images of other mismatches. This proportion of mirror image to orthogonal mismatches represents the minimum possible number of overlaps of this type under these conditions. It was considered desirable to sample the widest possible range of independent stimulus-response mispairings for the recognition tests. For all 16 mismatches to be independent would necessitate using a set of orthogonal Latin squares. Unfortunately, these do not exist for a six-by-six matrix (Winer, 1962).

In summary, two forms of the recognition test were constructed for each of the four groups of high- and low-meaningful stimuli from Lists 1 and 2, each test consisting of two halves. In a given half all stimulus and response items occurred only once, in two correctly matched and four mismatched pairs. The pairings in one half of the test were independent of the pairings of the items in the other half of the test. Thus each test contained four correct pairings and eight different mismatches of stimuli and responses. No stimulus-response pairings were repeated in the two forms of the test, although three pairs of mismatches occurred that were derived from the same pairs of correct pairings.

The final form of the recognition test, as used in the experiment, was constructed by arranging in random order the items from the first halves of the high- and low-meaningful recognition material, and then repeating this process for the second halves. This procedure was done separately for

Lists 1 and 2. The end product was four recognition tests, two for List 1 and two for List 2. Each test was in two parts, each part containing 12 stimulus-response pairings with both high- and low-meaningful stimulus terms in random order.

A practice list of eight nonsense shapes paired with eight one-digit-number responses was prepared. Three orders of the stimulus-response pairings for presentation, two orders of the stimuli only for recall testing, and one recognition test containing four correct pairs and four mismatches were used. This practice list was used to ensure that the subjects understood the instructions concerning the experimental procedure.

Four forms for rating the similarity of the stimulus items were composed, corresponding to the four combinations of high- and low-meaningful items from List 1 and List 2 (for an example, see Appendix A). Each form contained the 20 possible combinations of the six stimuli taken three at a time. The sequence of the triads was randomized separately for each form. Each triad of stimuli was arranged in a triangle, with one term representing the apex and the other two terms forming the base. The designation of the three terms into apex, left base and right base positions was randomized in each triad with the restriction that no stimulus term was repeated in a given position less than three times or more than four times.

Design

Each of 64 subjects went through three basic steps:

(a) They rated for similarity two sets of six stimuli.

(b) They attempted to learn 12 paired-associates for a prearranged number of trials by the study-test method (alternating presentations of the 12 pairs with instructions to learn the pairings, and presentations of the 12 stimuli only, with instructions to recall the missing responses). On the

last trial a recognition test was substituted for the recall test.

(c) They rated for similarity the two sets of stimuli previously rated, plus a third set of stimuli.

Differences in this procedure were as follows:

(1) Half of the 64 subjects rated for similarity the low-meaningful stimuli of both List 1 and List 2, the other half rated the high-meaningful stimuli of Lists 1 and 2. Half of the subjects did the List 1 ratings first and then rated the List 2 material while the other half rated the List 2 material first and List 1 second.

(2) Four groups of 16 subjects each attempted to learn the 12 word-pairs for either 2, 4, 8, or 16 trials, respectively. Half of the subjects in each group practiced on List 1, the other half on List 2.

(3) All subjects repeated the two similarity ratings they had done in (1), except that the order of the ratings was reversed. All subjects also rated for similarity the set of six stimuli from the list that they had attempted to learn responses to in (2), but which had not been rated previously, e.g., a subject who had first rated the high-meaningful stimuli from both Lists 1 and 2, and then had attempted to learn the stimulus-response pairings of List 1 would subsequently repeat his ratings on the high-meaningful stimuli from Lists 1 and 2, and in addition would rate the low-meaningful stimuli from List 1. Half the subjects repeated their previous ratings first before rating the new material, while the other half did the new rating before the old ones.

Table III summarizes the assignment of the combinations of the variables of list, meaningfulness, order of rating of the sets of stimuli, and the list used in the learning task for 16 subjects. This matrix of experimental conditions assigned to subjects was repeated four times, once

TABLE III

ASSIGNMENT OF SUBJECTS TO COMBINATIONS OF MEANINGFULNESS AND LIST MEMBERSHIP OF STIMULI IN SIMILARITY RATINGS AND OF LIST USED IN LEARNING TASK

SUBJECT	MEANINGFULNESS & LIST OF STIMULI IN PRELEARNING RATINGS		LIST USED IN LEARNING TASK AND RECOGNITION TEST	MEANINGFULNESS & LIST OF STIMULI IN POSTLEARNING RATINGS		
	ORDER RATED			ORDER RATED		
	1st	2nd		1st	2nd	3rd
1	H1	H2	1	H2	H1	L1
2	H1	H2	1	L1	H2	H1
3	H1	H2	2	H2	H1	L2
4	H1	H2	2	L2	H2	H1
5	H2	H1	1	H1	H2	L1
6	H2	H1	1	L1	H1	H2
7	H2	H1	2	H1	H2	L2
8	H2	H1	2	L2	H1	H2
9	L1	L2	1	L2	L1	H1
10	L1	L2	1	H1	L2	L1
11	L1	L2	2	L2	L1	H2
12	L1	L2	2	H2	L2	L1
13	L2	L1	1	L1	L2	H1
14	L2	L1	1	H1	L1	L2
15	L2	L1	2	L1	L2	H2
16	L2	L1	2	H2	L1	L2

for each group of subjects corresponding to the four degrees of learning used in the experiment.

Apart from the fact that the two ratings repeated after the learning task always occurred adjacent to each other and in the reverse order that they were assigned for the initial rating, the variables described in (1), (2), and (3) above are orthogonal to each other. As this constitutes a 2 x 2 x 4 x 2 x 2 factorial design, containing 64 cells, it can be seen that each subject represents a unique combination of these variables. However, as it seems highly unlikely that there would be a significant interaction between certain combinations of variables such as the order in which the subjects did the post-learning similarity ratings and the list which they attempted to learn, any meaningful analysis of variance done on the data in this design would have at least four subjects per cell.

A number of other variables were controlled, but not in an orthogonal design.

(1) Within each of eight groups of eight subjects (each group being defined by the eight combinations of the two lists used in learning and the four numbers of trials for which practice continued), the two forms of the recognition test were used equally often. (Two exceptions occurred to this rule through an error by the experimenter--in two groups, one recognition test was used five times, and the other three times).

(2) Within each of the eight lists-by-trials treatment groups the two orders of administration of the separate halves of the recognition test occurred equally often. (Three exceptions to this rule were caused by experimenter's error. In three groups, one sequence was used five times. In all, four subjects were affected by these errors and those mentioned in (1) above as two errors in procedure were committed in the running of one

subject.)

(3) Except for the previously mentioned errors, all four possible combinations of the two forms of the recognition test and the two orders of presentation of the halves were used equally often and randomly assigned to subjects in each of the eight treatment-combination groups.

For a number of reasons it is felt that these errors in procedure (which were discovered only after the conclusion of the experiment) were sufficiently minor that their possible effects on the data can be ignored. First, while the variables involved (groups of specific stimulus-response mispairings and sequence effects) could quite conceivably affect the data of the experiment, they are of little theoretical interest in this study and were controlled only so that their effects were not confounded with those of other more important variables. Second, it was considered that the disruptive effects of these irregularities would be minimal, as the data of only four subjects out of 64 were affected, and the smallest group that contained two of these subjects also contained 14 others.

The particular treatment combinations were randomly ordered in advance, and assigned consecutively to subjects as they arrived at the laboratory.

Subjects

The subjects were 64 University of Alberta students (33 males and 31 females) enrolled in the introductory psychology class who took part in the experiment to fulfill a course requirement. The majority of subjects had previously served in other verbal learning experiments using a variety of material and procedures, none of which particularly resembled those used in the present study.

Nine other subjects had originally taken part in the experiment,

but were replaced because of equipment breakdown, experimenter error, or subsequent discovery that they had not followed instructions in the similarity rating task. Their data were replaced by testing subsequent subjects in their place. One subject found the learning task upsetting and was permitted to leave the experiment.

Apparatus

The stimuli and responses were presented visually to each subject with two "One-plane Readout" display cells manufactured by Industrial Electronic Engineers Inc. The subject sat in a cubicle with the display cell screens mounted flush in the wall at eye level about 2 feet away. The letters making up the stimuli and responses were projected as white against a dark background, and were approximately 1 inch in height. The selection of stimuli and responses was programmed with a Western Union tape reader and a bank of relays. The timing was controlled by a synchronous motor and an eccentric cam.

Procedure

Similarity Ratings. Before the initial rating, the subject was told to read a dittoed sheet describing the rating procedure (Appendix B). He was instructed to indicate the most similar and least similar pairs of stimuli in each triad, with the criterion for similarity to be decided by him. He was then given his two initial preassigned similarity rating sheets.

For the post-learning similarity rating, the subject was merely reminded to follow the same procedure as before when he was given his three rating sheets. Most subjects took approximately 5 minutes to complete the ratings on one set of stimuli.

Learning Task and Recognition Test. After finishing the initial similarity ratings the subject was seated in front of the display cells

while the instructions (Appendix C) were read to him by the experimenter from outside the cubicle. The nature and timing of the study-test method was outlined to him, and he was instructed to call out as many responses as he could remember on test trials. He also received instructions concerning how to respond on the recognition test, and was told that this type of test would be substituted for a recall trial at some point during the learning task.

After any questions had been answered, the subject was told that he would be given a practice run on the procedure using pairs of shapes and numbers. He was then shown two presentations of the eight shape-number practice pairs, each presentation being followed by a test presentation of the eight stimuli alone. The pairs were presented for 2 seconds each, and the stimuli for 4 seconds each. There was a 6-second interval during the transition from presentation to test, and from test to presentation. If the subject did not attempt a response during the first recall test trial, he was reminded that he should try to call out any responses that he could remember. All subjects attempted a number of responses by the second recall trial.

A third presentation of the pairs was made 6 seconds after the second recall test. The subject was then given reminder instructions (Appendix C) on the recognition test, and shown four of the correct pairings and four mispairings of the stimuli and responses in a random sequence. He was given unlimited time to respond to each pair. As all subjects showed appreciable learning at this point in the procedure, it was possible to detect the few subjects who had apparently misunderstood the instructions about the probability estimates and consistently reversed the procedure for reporting their estimates of the probability of a change (i.e., responded

"zero" instead of "one hundred" when they were certain the pair had been changed). These subjects were asked to repeat the instructions and were reminded if they were incorrect.

The subject was then given the preassigned number of presentation trials on the preassigned list. Each presentation except the last was followed by a recall test trial. On the presentation trials, the 12 pairs were shown for 2 seconds each; on the test trials, the 12 stimuli were shown for 4 seconds each. All responses and omissions to stimuli were recorded, including corrections by the subject of an initial response or responses. A 6 second-pause was interposed between presentation and test trials.

After the final presentation trial, the subject was read the same "reminder" instructions concerning the recognition procedure as were used in the practice procedure. Approximately 15 seconds elapsed between the end of the last presentation trial and the beginning of the recognition test, while the experimenter read the instructions and adjusted the apparatus. The subject was then shown the 24 pairs of the recognition test, each pair being shown until the subject responded. The subject's final response to each pair was recorded. After the recognition test was completed, the subject was told to rate for similarity the last three sets of stimuli, using the same procedure as before, and was then dismissed.

Treatment of Similarity Rating Data. The data were first converted from their triadic form to paired rankings, and from these the best complete order of ranking of the 15 pairs of stimuli from each set was determined. The criteria by which a given order was decided to be best are discussed in Chapter III. The procedure for determining the best order was as follows: (1) The method of triangular analysis (Coombs, 1964) was used to determine the best partial order (including ties between pairs which had not been

compared directly) of the pairs which could be derived from the data. As this involved permuting columns and rows of a 15 x 15 matrix, a computer program was used which first provided a rough solution by ranking the sums of entries in the rows and then reduced the labor of doing the remaining adjustments manually.

(2) The best complete order of the pairs was then determined manually from the partial order in (1) by applying the criteria discussed in the next chapter.

CHAPTER III

Results

Analysis of Similarity Rating Data

The first step in the analysis of the data is the derivation of a set of ranked interstimulus distances from each subject's responses on the similarity rating tasks. There exist a number of possible procedures for accomplishing this purpose which correspond to given data reduction models or to assumptions concerning the processes underlying the pattern of response. Before the procedure used in this study is described, a brief review of the objectives of the experiment will be helpful for an understanding of the rationale behind the techniques used.

The purpose of this study is to evaluate a procedure for measuring the degree of similarity between a number of verbal stimuli. It was decided to develop a technique for producing an ordinal scale, i.e., a rank-order of interstimulus distances, for each subject.

A major task in the evaluation of this technique is to demonstrate that it can produce results of satisfactory reliability and validity. Measures of reliability and validity typically involve the calculation of coefficients of correlation between the test being studied and some criterion measure; in the present study, the most appropriate measure of correlation is the rank correlation coefficient, or Spearman's rho (Siegel, 1956). The calculation of Spearman's rho involves ranking a set of individuals on two variables. The smaller the differences between the two ranks each subject

receives for his scores on the two variables, the greater the value of rho, and the greater the relationship indicated between the two variables.

To determine the rank correlation coefficient it is not necessary that individuals be allotted different scores on each of the two variables. It is possible to calculate a correlation coefficient when a number of individuals have the same score on a variable (i.e. when the measure of the variable does not discriminate between the amounts of the variable that several subjects possess). The subjects with equal scores are allotted the same rank, which is the median of the ranks that the group would have had if they had been differentiated. The coefficient of correlation is then calculated as before.

However, as the calculation of Spearman's rho is an adaptation of the Pearson correlation coefficient based on the assumption that the intervals between ranks are equal, the estimate of the relationship between two variables through rho will be distorted if too many tied ranks are allowed to occur. It is obvious, then, that any demonstration of statistically significant reliability and validity of the measurement procedure under study must avoid the possibility that this significant relationship is trivial. In the interests of rigorously evaluating the procedure being studied, therefore, an attempt was made to minimize the number of tied ranks in the ordinal inter-stimulus distance scale derived from the similarity rating data.

This was done by applying successive assumptions to the data, each assumption in the series yielding information not provided by the preceding assumption. However, the certainty with which the information yielded by a given assumption was regarded decreased with the assumption's rank in the series. To put this in another way, a number of different models were

used to interpret the data. These models are rank ordered so that if they yield conflicting information, the results of a higher-ranked model would be accepted over those of a lower-ranked model. In general, each model yields some unique information and some information shared with that yielded by other models (where these several sources of information might be in agreement, or conflicting). The plan is to derive a rank-ordering of the interstimulus distances by first using all of the information provided by the model in which the most confidence can be placed, then using the information from the next-ranked model which was not supplied by the first model, and so on.

The first assumption is that the rank-ordering of the inter-stimulus distances for a given set of similarity rating data is transitive. From this it follows that the dominance relations between stimulus pairs must be transitive, i.e., if $AB > BC$, and $BC > AC$, then $AB > AC$, where XY signifies "the distance between stimuli X and Y". In the case where the data indicate that $\overline{AB} > \overline{BC}$ and $\overline{BC} > \overline{CD}$, but the distances $\overline{AB}$ and $\overline{CD}$ have not been compared in the experimental procedure, by the first assumption we can conclude that $\overline{AB} > \overline{CD}$.

However, it is sometimes the case that three pairs of distances have been rated in the data but have indicated an intransitive relationship. For example, a subject may have responded that $\overline{AB} > \overline{BC}$, $\overline{BC} > \overline{AC}$, and $\overline{AC} > \overline{AB}$. In this case, it is assumed that this intransitive series is due to an error in rating by the subject. Where one or more intransitivities occur in a subject's similarity ratings the most probable true rank-order of the inter-stimulus distances is assumed to be the one which contradicts the data least, i.e., that would require the fewest alterations to the data to produce. If more than one ranking of the distances can be achieved with the same minimal

number of data changes, each distance is allocated the mean of the ranks that it receives in the different orderings.

Frequently sets of partial rankings such as

$$\overline{AB} > \overline{BC} > \overline{CD} > \overline{DE}$$

and $\overline{AB} > \overline{AE} > \overline{DE}$

are found, where $\overline{AE}$ had not been compared with $\overline{BC}$ or $\overline{CD}$. The position of $\overline{AE}$ is indeterminate relative to $\overline{BC}$ or $\overline{CD}$ by the first assumption. The second assumption, which covers cases such as this, is that:

$$\text{if } \overline{AB} > \overline{AE} > \overline{DE},$$

$$\text{then } E(\overline{AB} - \overline{AE}) = E(\overline{AE} - \overline{DE}),$$

$$\text{and if } \overline{AB} > \overline{BC} > \overline{CD} > \overline{DE},$$

$$\text{then } E(\overline{AE} - \overline{BC}) = E(\overline{BC} - \overline{CD}) = E(\overline{CD} - \overline{DE}),$$

where $E(X_i)$ signifies the expected value of X_i in the population X. On the assumption that, in general, $E(UV + WX) = E(UV) + E(WX)$ from the first equation it can be seen that

$$E(\overline{AE}) = E(\overline{AB} + \overline{DE})/2$$

and from the second equation

$$E(\overline{BC}) = E(\overline{AB} + \overline{CD})/2$$

$$\text{and } E(\overline{CD}) = E(\overline{BC} + \overline{DE})/2$$

$$\text{since } \overline{CD} > \overline{DE},$$

$$\text{therefore } E(\overline{BC}) > E(\overline{AE}),$$

$$\text{and since } \overline{AB} > \overline{BC},$$

$$\text{therefore } E(\overline{AE}) > E(\overline{CD}).$$

From the above it can be seen that the most probable ranking of these five distances, according to the second assumption, is $AB > BC > AE > CD > DE$. It is thus possible to produce a mutual ordering of the intermediate distances in two unequal series of distances when the first and last members

of the series are the same for the two series.

The third assumption can sometimes produce a ranking not given by the first or second assumptions, where the data give the following cases:

(1) $\overline{AB} > \overline{BC} > \overline{CD}$, and $\overline{AB} > \overline{AD} > \overline{CD}$, or

(2) $\overline{AB} > \overline{BC}$, $\overline{BC} > \overline{AC}$, and $\overline{AC} > \overline{AE}$

In the first case, the non-compared distances $\overline{BC}$ and $\overline{AD}$ are intermediate between $\overline{AB}$ and $\overline{CD}$, but the ranking cannot be resolved by invoking Assumption 2 as the series are equal in length. The second case produces three equally probable rankings, according to Assumption 1: $\overline{AB} > \overline{BC} > \overline{AC}$, $\overline{AC} > \overline{AB} > \overline{BC}$, and $\overline{BC} > \overline{AC} > \overline{AB}$.

Assumption 3 states that the rank order of a given inter-stimulus distance is a function of the number of other distances that it surpasses. Thus, to resolve the indeterminate relative rankings of $\overline{AD}$ and $\overline{BC}$ in Case 1 above, and of $\overline{AB}$, $\overline{BC}$, and $\overline{AC}$ in Case 2, highest ranking is given to the distance that dominates the largest proportion of all the distances with which it was compared in the similarity ratings, second highest rank is given to the distance that had the next largest sum of "greater" judgments, and so on.

In summary, the procedure for determining the rank order of the interstimulus distances from the similarity ratings runs as follows:

(1) The distances were arranged in the complete rank order that required the fewest number of changes in the subject's pair-wise similarity ratings. Occasionally this was all that was required for a given set of data; each distance had been compared in the similarity ratings with the distances immediately above and below it. More often, two or more series of distances derived from a set of ratings would begin and end with common members, but the intermediate members of the chains had not been compared.

(2) When two series, containing x and y stimulus pairs, respectively, began and ended with common members, where $x \neq y$, each of the x - 1 differences between succeeding interstimulus distances in the first series was set as equal to $\frac{1}{x - 1}$, and the y - 1 differences between adjacent pairs in the second series were each set equal to $\frac{1}{y - 1}$. The two series were then merged into a common order.

Occasionally it was found that three or more series occurred in a given set of data having different initial and final members. This created a problem, as different common orders could be found depending on which two series were picked to be merged first. It was decided that in cases such as this the longest partial order including the greatest and least distances in the set would be used as a standard, and that all shorter partial orders would be subordinated to this.

(3) In cases where there were two non-compared pairs tied for the same ranks, including cases of parallel series each containing an identical number of members, the ties were broken by allotting the higher rank to the pair which dominated the higher proportion of all the pairs to which it had been compared. This strategem was also used to resolve ties caused by circular chains due to intransitivities; the highest rank went to the pair in the chain having the highest total "vote count", and the remaining pairs were ordered according to the pair-wise ranking of the similarity estimation data.

It can be noted that Assumption 1 is what Coombs (1964) calls a "decomposition model". He recommends it for the study of similarities data gathered through the present method because it uses only that information concerning the relative similarity of a pair of distances that was derived from a subject's ratings of the two distances together. Assumption 2 has

not, to the writer's knowledge, been used in this type of scaling before; its chief virtue is its parsimony in postulating that the differences between a given interstimulus distance and the distances immediately above and below it are equal. Whether this assumption is valid is a question of fact, and would seem best studied empirically. The third assumption corresponds to what Coombs (1964) refers to as an "expected matrix model"; he does not recommend its use for the present method of gathering similarities data because it utilizes information about a pair of distances derived from comparisons of the single distances with different groups of other distances. This, he feels, distorts the rankings obtained because of greater context effects and the greater instability of differing comparisons. However, this assumption was felt to be useful because of the information it provided on non-compared pairs.

No other justification of these assumptions will be presented at this time, apart from mentioning that they appear to have achieved their intended purpose of minimizing ties in the distance rankings. Out of 320 rankings of 15 interstimulus distances there occurred only 285 ties involving two distances, 55 ties involving three distances, and 14 ties involving four distances. As a two-way tie means that the ranking of one pair of distances is indeterminate, a three-way tie that the rankings of three pairs are indeterminate, and a four-way tie renders indeterminate six paired rankings, it can be seen that the ranking of only 534 pairs of distances was unavailable from the above procedure. This represents less than 1% of the total of 67,200 rankings of pairs of distances implicit in the data of this experiment. Further scrutiny of the assumptions is deferred not because it is felt that the procedure is free from major criticism, but because it is felt that this discussion is best postponed until these assumptions can be

studied in the light of relevant empirical evidence. The purpose of the present study is to validate the procedure by demonstrating that similarity ratings are related to generalization errors in paired-associate learning.

Reliability of Similarity Rankings

Internal Consistency. After the five sets of distance rankings had been established for each subject, the data were examined with regard to corrections for intransitivities in the rankings. As the procedure assumes that the pair-wise comparisons can lead to a transitive ordering of the interstimulus distances, whether or not the intransitivities found can be ascribed to chance error in the similarities data is of interest in verifying this assumption. If the average number of intransitivities were larger than could be reasonably attributed to measurement error, then the validity of the basis of entire procedure would be in serious doubt.

It was found that the mean number of intransitivities per test over all tests on all subjects was 2.92. Little is known about the characteristics of the expected distribution of intransitivities in this data-collection method, so no test of significance can be applied to these data. However, as this figure indicates that of the 60 paired comparisons made by each subject, an average of less than 5% were inconsistent with each other, it would seem that this finding supports the assumption that the underlying similarity relationship for the stimuli studied is essentially transitive.

The occurrence of intransitivities is of interest because it gives an index of the reliability of the similarity rankings (Vastenhouw, 1962), and can be considered as a measure of the internal consistency of the measuring instrument. The intransitivities data were examined to determine whether frequency of intransitivities is relatively stable over all conditions, or whether subjects behave more inconsistently under some

circumstances than under others. The effects on frequency of intransitivities of the variables of previous experience in ranking for similarity, meaningfulness of material, specific list material, and experience with the material as paired-associate stimuli were studied to evaluate the data's potential for further analysis.

It will be recalled that each subject rated for similarity five sets of six stimuli. Six of the stimuli in the pairs to be learned and six control stimuli were rated before the learning task, and then the ratings were repeated after learning. The material in these ratings was either all high-meaningful or all low-meaningful for a given subject. In addition, each subject rated the second set of experimental stimuli (of opposite meaningfulness to the set previously rated) after the learning task, this material, of course, being rated for the first time.

An inspection of the data suggested that this last-mentioned material had a markedly greater number of intransitivities associated with it than did the other sets (see Table IV). An analysis of variance over the five ratings showed a significant F - ratio for differences between treatments, $F(4, 252) = 10.05$, $p < .005$. When the mean number of intransitivities for the experimental material rated for the first time after learning is compared to the other four means combined, an F-ratio of 39.46 is found.

TABLE IV

MEAN NUMBER OF INTRANSITIVITIES IN FIVE SIMILARITY RANKINGS OF STIMULI

	Before Experimental	Control	Experimental Repeated	Control Repeated	Exper.non-Repeated	Total
Mean	2.63	2.45	2.66	2.72	4.23	2.92

Because this is an *a posteriori* test, Scheffé's (Winer, 1964) procedure is used to determine the criterion for significance. The value of *F* necessary to reach significance at the .005 level of probability for 4 and 252 degrees of freedom by this procedure is about 24.0; thus the observed difference is clearly significant. No other differences between treatment means were significant. It may be noted that the F-ratio (F (63, 256) = 2.338) for between-subjects effects was significant at the .001 level, indicating that there are reliable individual differences in frequency of intransitivities.

A possible explanation for the difference between treatment means is that the subjects became set in the dimensions of similarity that they attended to after the first two or four similarity ratings, which were all of the same level of meaningfulness. Consequently, when asked to rate material of a different level of meaningfulness, they found it difficult to adjust to different criteria of similarity, and as a result showed an increase in the inconsistency of their ratings.

It was next decided to determine if there were differences in inconsistency of response as a function of the meaningfulness of the material and the two lists of stimuli used. This analysis was done on only the material rated before the learning task, as it was wished at this point to determine the effects of meaningfulness and list without the added complication of whether or not the material had occurred as stimuli in a learning situation, or whether the material was being scaled for the first or second time. A 2 x 2 factorial analysis of variance, with repeated measures on one variable was carried out. The means appear in Table V and a summary of the analysis in Table VI.

TABLE V

MEAN NUMBER OF INTRANSITIVITIES IN PRE-LEARNING SIMILARITY RANKINGS AS A FUNCTION OF MEANINGFULNESS AND LIST.

	List	1	2	Mean
Meaningfulness	High	2.91	2.31	2.61
	Low	2.09	2.84	2.48
	Mean	2.50	2.58	2.54

TABLE VI

SUMMARY OF ANALYSIS OF VARIANCE OF NUMBER OF INTRANSITIVITIES IN PRE-LEARNING SIMILARITY RANKINGS

Source	df	SS	MS	F
Between subjects				
Meaningfulness (M)	1	0.633	0.633	< 1
Subjects within groups	62	221.672	3.575	-
Within Subjects				
Lists (L)	1	0.196	0.196	< 1
M x L	1	14.445	14.445	6.74*
L x Subjects within groups	62	132.859	2.143	-

* Significant at the .05 level of probability.

It can be seen that the only significant effect is the interaction between lists and levels of meaningfulness. Table V shows that the high-meaningful material in List 1 produced more inconsistent response than did the low-meaningful material, whereas the converse was true of List 2. As it will be recalled that an attempt was made to design the two lists to be as equivalent as possible, it would seem that there are potent material-

specific effects on consistency of rating that cannot be predicted from general properties such as meaningfulness.

The final statistical evaluation of the intransitivities data concerned the combined effects of meaningfulness, list, and degree of exposure to the material as stimuli in a paired-associate learning task on the stability of consistent responding. Changes in frequency of intransitivities between before and after the use of the rated material as stimuli in 2, 4, 8, and 16 trials of paired-associate learning were studied in an attempt to determine if distance-ranking data measures taken at various points in a learning experiment could be considered equally reliable.

This was done by an analysis of variance on the two differences (in the experimental and control lists) between number of intransitivities before and after the learning task for each subject. The data were transformed to eliminate negative values by adding a constant factor of 8 to each score.

It can be seen from Table VII that only the effect due to the interaction between levels of meaningfulness and trials is significant. This relationship is depicted graphically in Fig. 1, which shows that for the high-meaningful material the frequency of intransitivities decreased slightly after 2, 4, and 16 trials of learning, but showed a sharp increase after 8 trials, while the curve for low-meaningful material is U-shaped, being highest for 2 and 16 trials, and lowest for 4 and 8 trials. No interpretation can be suggested for this configuration. Considering that the statistical significance of this difference was marginal, that it was one of 15 independent tests in the analysis of variance, and that no meaningful interpretation seems apparent, further discussion of this effect would seem fruitless.

TABLE VII

SUMMARY OF ANALYSIS OF VARIANCE OF CHANGE IN NUMBER OF INTRANSITIVITIES IN SIMILARITY RANKINGS BETWEEN BEFORE AND AFTER PAIRED-ASSOCIATE LEARNING

Source	df	SS	MS	F
Between subjects				
Meaningfulness (M)	1	0.945	0.945	< 1
Trials (T)	3	12.710	4.237	1.228
M X T	3	37.960	12.653	3.668*
E X L	1	0.070	0.070	< 1
E X L M	1	11.882	11.882	3.444
E X L X T	3	9.460	3.153	< 1
E X L X M X T	3	14.898	4.966	1.439
Subjects within groups	48	165.625	3.450	-
Within subjects				
Experimental condition (E)	1	0.382	0.382	< 1
E X M	1	5.695	5.695	< 1
E X T	3	3.523	1.174	< 1
E X M X T	3	13.460	4.487	< 1
List (L)	1	2.820	2.820	< 1
L X M	1	1.320	1.320	< 1
L X T	3	6.710	2.237	< 1
L X M X T	3	2.460	0.820	< 1
Residual (within)	48	320.125	6.669	-
Total	127	-	-	-

* Significant at the .05 level of probability.

To summarize the findings with respect to internal consistency of the similarity measure as indicated by frequency of intransitivities, it was found that subjects' rating showed a high level of consistency overall. The data seemed to show that subjects became more inconsistent in their ratings when they were given a set of stimuli whose similarity was higher or lower than the previous sets that they had rated. Significant differences in rating consistency were found as a function of specific sets of stimuli that did not seem predictable from the degree of meaningfulness of

the stimuli nor could they be controlled by constructing lists according to apparently similar principles.

Test-Retest Reliability. The next step in the evaluation of the similarity rankings was an examination of their test-retest reliability. This was done by first calculating test-retest correlation coefficients between the pre- and post-learning control stimulus rankings, and then studying the stability of reliability coefficients under various variables.

Table VIII shows the mean test-retest rank correlation coefficients over 16 subjects for the derived rankings of four control sets of stimuli defined by the combination of two lists and two levels of meaningfulness. The control sets of stimuli, it will be remembered, are the ones which were not used in the learning task; thus the operation used to determine these correlation coefficients was to have each subject rate for similarity a set of six stimuli, next attempt a learning task involving different stimuli, and then re-rate the stimuli.

TABLE VIII

MEAN RANK ORDER TEST-RETEST
RELIABILITY COEFFICIENTS FOR CONTROL STIMULI

		LIST		
		1	2	Mean
Meaningfulness	High	0.597	0.699	0.648
	Low	0.783	0.712	0.748
	Mean	0.690	0.706	0.698

It should be emphasized that these coefficients are a measure of the correlation between the two sets of rankings derived from the before and after ratings.

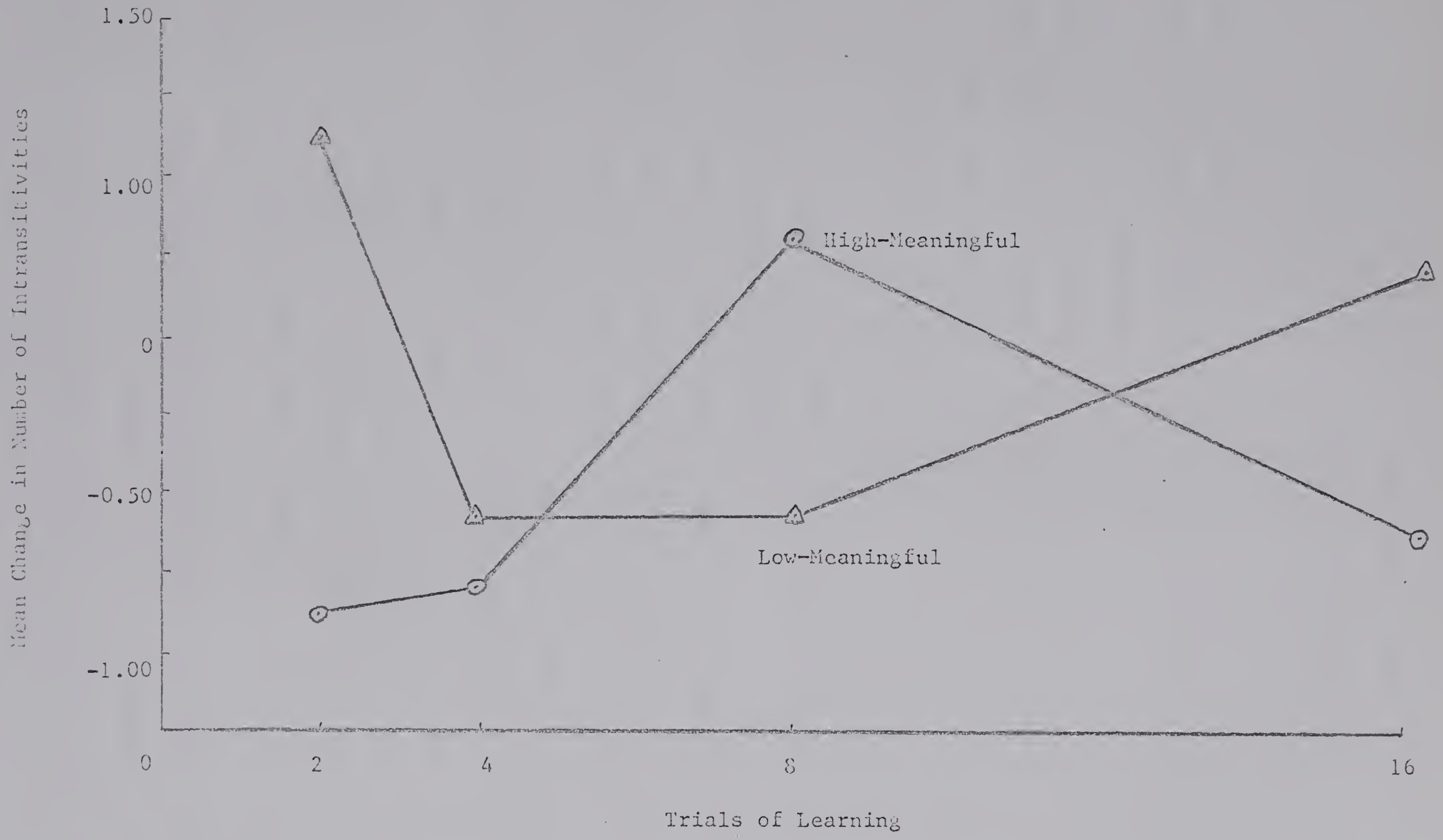


Figure 1. Mean change in number of intransitivities before and after Learning.

All the mean correlation coefficients in Table VIII, if reported for a single subject, would be significant at the .05 level of probability. Of the 16 subjects rating the high-meaningful stimuli, 11 had reliability coefficients significant at a maximum of $p < .05$; for the List 2 high-meaningful material subjects, 13 out of 16 coefficients were at less than the .05 level of probability; for the ratings of the List 1 low-meaningful stimuli, 15 out of 16 coefficients were significant at less than the .05 level; while for the List 2 low-meaningful stimuli 14 out of 16 reliability coefficients were significant at or beyond the .05 level of probability. Of the 64 individual reliability coefficients calculated, 6 were significant with $.01 < p < .05$, 16 were significant with $.001 < p < .01$, and 30 with $p < .001$, while only 12 coefficients were not significant at equal to or less than the .05 level of probability.

Considering the abbreviated nature of the measuring instrument and the fact that the reliability coefficients express the consistency of not only the subjects' rating behavior but also of the procedure that was used to derive the rankings of the distances, the observed mean reliability coefficient of 0.698 would seem to indicate a satisfactory level of stability. It is interesting to note that the average reliability seems to be higher for the low-meaningful stimuli than for the high-meaningful material. Also, within levels of meaningfulness, the higher correlation coefficients are associated with the lower mean frequencies of inconsistencies (Table V), and vice versa.

Having established that the procedure under study is moderately reliable, the effects of the variables of list, meaningfulness, use of the stimuli as experimental or control material in a learning task, and number of trials of attempted learning upon the consistency of subjects' ranking

of the interstimulus distances were studied.

This was done through an analysis of variance on the reliability coefficients of the rankings of the control and experimental stimuli. The data were transformed by first converting the values of the correlation coefficients to Fisher's Z_r to normalize the sampling distribution, and then adding 1 to all scores to eliminate minus values. A summary of the analysis of variance is presented in Table IX. It can be seen from this summary that only two effects were statistically significant: levels of meaningfulness, and the four-way interaction between experimental condition, list, meaningfulness and trials.

A graph of the four-way interaction is shown in Figure 2. It would seem difficult to provide a meaningful interpretation of this apparently random relationship. When it is considered that the estimate of the effects of the four-way interaction is based on a between-groups comparison, each point on the graph being determined by the data from only four subjects, that the statistical significance of this interaction was marginal, and that none of the four three-way interactions has an F-ratio greater than unity, no reasonable explanation for the magnitude of the four-way interaction seems apparent. It might be pointed out that the four-way interaction could be regarded as the interaction between the effect of the three-way interaction of experimental condition, meaningfulness, and trials (a theoretically meaningful interaction with a negligible mean square) and the effect of the factor of lists (a factor of little theoretical interest, as it represents essentially random item-specific differences).

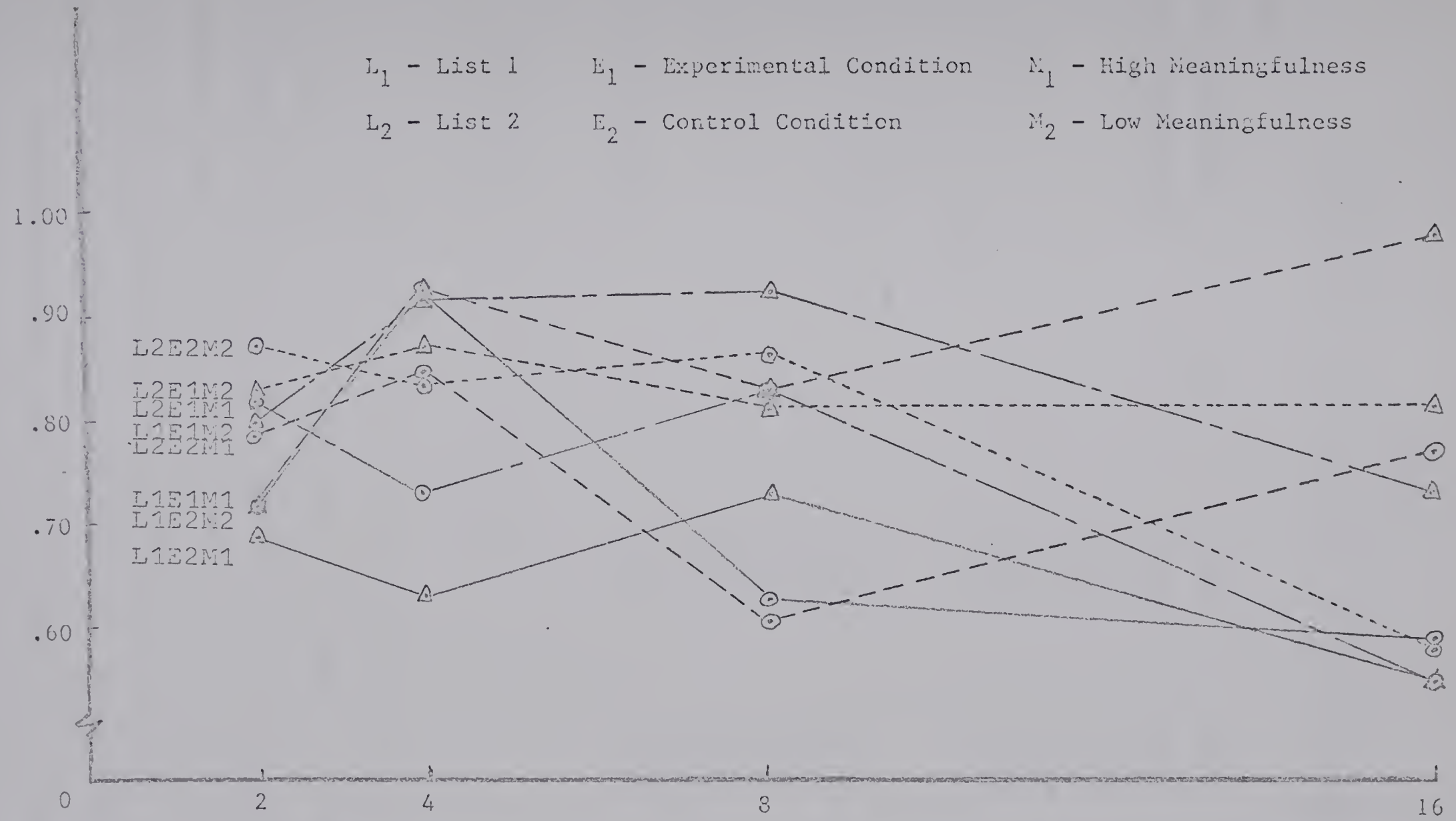


Figure 2. Mean correlation coefficients between before- and after-learning stimulus pair rankings as a function of list, experimental condition, meaningfulness, and trials.

TABLE IX

ANALYSIS OF VARIANCE OF TEST-RETEST RELIABILITY COEFFICIENTS FOR RANKED INTER-STIMULUS DISTANCES

Source	df	SS	MS	F
Between subjects				
Meaningfulness (M)	1	2.732	2.732	8.229**
Trials (T)	3	1.600	0.533	1.777
M X T	3	0.347	0.116	<1
E X L	1	0.004	0.004	<1
E X L X M	1	0.278	0.278	<1
E X L X T	3	0.727	0.242	<1
E X L X M X T	3	2.890	0.963	2.901*
Subjects within groups	48	15.956	0.332	-
Within subjects				
Experimental condition (E)	1	0.040	0.040	<1
E X M	1	0.054	0.054	<1
E X T	3	0.516	0.172	1.186
E X M X T	3	0.040	0.013	<1
List (L)	1	0.030	0.030	<1
L X M	1	0.571	0.571	3.938
L X T	3	0.915	0.305	2.103
L X M X T	3	0.417	0.139	<1
Residual (within)	48	6.980	0.145	-
Total	127	-	-	-

* Significant at .05 level of probability

** Significant at .01 level of probability

Concerning the significant meaningfulness effect, the mean reliability coefficient for the high-meaningful material is 0.665 and for the low-meaningful material is 0.768. The difference between these figures is approximately equal to that reported earlier for the control material alone. Thus, it can be concluded that the reliability of similarity rankings is greater for the low-meaningful material than for the high-meaningful stimuli.

The hypothesis that extensive experience with stimuli in a paired-associate learning task might change the similarity relations between the

stimuli was considered next. If this hypothesis were correct, it would be expected that the test-retest reliability of the experimental material would decrease as a function of trials, while the reliability of the control stimuli would remain relatively unchanged. If this occurred, it should be evidenced in the analysis of variance as a significant experimental conditions by trials interaction. No such significant interaction is apparent in Table IX.

Fig. 3 shows a graphical representation of the relevant data. It can be seen that the difference in reliability between experimental conditions is relatively constant for the groups tested after two, four, and eight trials, but shows the predicted drop for the experimental condition on Trial 16. It is possible that a large number of learning trials must occur before the hypothesized drop in reliability is found, and this drop had been obscured by having three tests during the first eight trials of learning but only one test on the sixteenth trial. If this were the case, then the correct test of the hypothesis would be to compare the difference between control and experimental material of Trial 16 with the combined experimental-control differences on Trials 2, 4, and 8. However, the F-ratio for this comparison was only 3.21 with 1 and 48 degrees of freedom. As an F-ratio of 4.04 is necessary for significance at the .05 level with these degrees of freedom, the hypothesis that there is no difference in reliability between experimental and control material for differing degrees of practice on a paired-associate learning task cannot be rejected on the evidence presented here.

Degree of Concordance Among Subjects. It would be of interest to know if all subjects gave basically the same rankings when rating each set of stimuli, or if there are individual differences in the rankings of

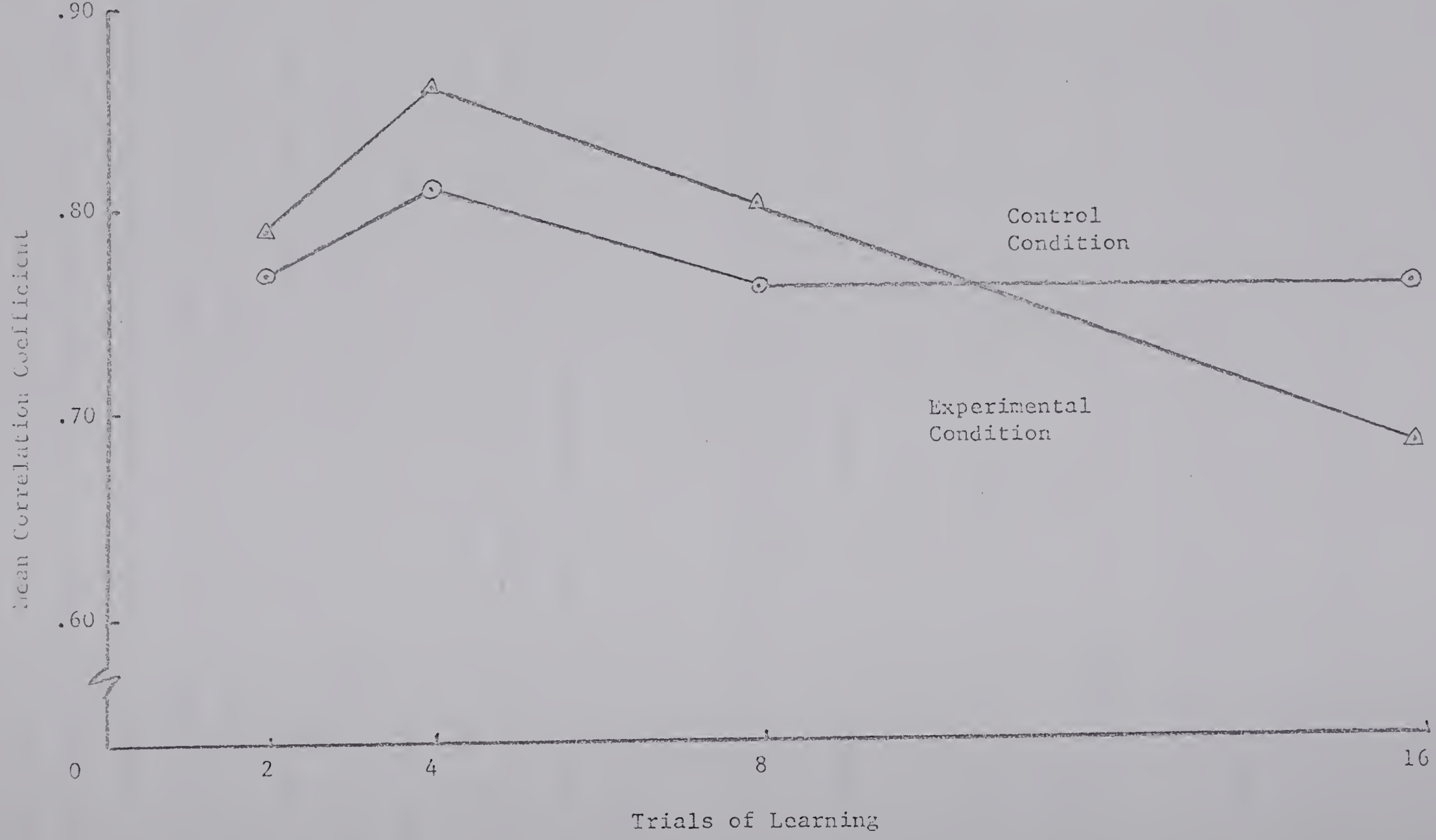


Figure 3. Mean correlation coefficients between before- and after-learning stimulus pair rankings as a function of experimental condition and trials.

different subjects. It had been hypothesized earlier in this paper that there are appreciable differences in inter-subject consistency in ranking the interstimulus distances for highly meaningful stimuli. To measure the amount of agreement between subjects on the rankings of the four sets of stimuli, W, Kendall's coefficient of concordance (Siegel, 1956) was calculated for these data (Table X). All reported values of W are significant at the .001 level of probability. The coefficient of concordance is a measure of the amount of agreement between a number of subjects on the ranking of a set of stimuli (here, distances). "Concordance" is a function of the average rank correlation coefficient that would be found if this were calculated for all possible pairs of rankings; Table X also shows this average value of rho, the rank correlation coefficient for the four sets of stimuli.

TABLE X

COEFFICIENTS OF CONCORDANCE (W) AND AVERAGE RANK CORRELATION (RHO_{AV}) FOR FOUR SETS OF STIMULI

Meaningfulness	High		Low	
List	1	2	1	2
Coefficient of Concordance (W)	.303	.346	.718	.681
Average Rank Correlation Coefficient (rho_{AV})	.281	.325	.709	.671

It will be noticed that the average correlation coefficients are quite small for the high-meaningful material, but for the low-meaning stimuli they are only slightly less than the mean reliability coefficients (Table VIII) for the low-meaningful material. Also, the rankings of W and

rho_{AV} are the same within each level of meaningfulness as the rankings of the reliability coefficients.

From these results it would seem that there was some communality among subjects' rankings for all four sets of stimuli. However, there would seem to be more stereotypy of response for the low-meaningful stimuli than for the high-meaningful stimuli. In fact, there is reason to believe that virtually all of the measured deviation from perfect agreement among subjects about the ranking of the low-meaningful material is due to errors of measurement, when it is considered that the average correlation coefficient for the List 1 distance rankings is only .065 less than the reliability coefficient, while the difference for the List 2 material is only .027. The corresponding differences for the high-meaningful material, in contrast, are .316 and .374 for Lists 1 and 2, respectively. If a correction for attenuation is made to estimate what the average correlation between subjects would be with perfectly reliable tests, the corrected rhos for the low-meaningful material are .905 and .942, while the average correlations for the high-meaningful material are only .470 and .464. This would seem to indicate that there was a considerable amount of disagreement between subjects as to the ranking of the distances between the high-meaningful stimuli, although individual subjects showed reasonable consistency in repeating their rankings.

The mean ranking given to each interstimulus distance was calculated for the low-meaningful material of List 1 and List 2; the stimulus pairs and the means are shown in Table XI. The pairs in Lists 1 and 2 have been matched in Table XI according to their common structures of the letter elements of which they are composed. A Pearson product-moment correlation coefficient of 0.910 (significant at the .001 level of probability for 13 df)

was calculated for the relationship between the two sets of mean ranks, showing that the subjects showed a high communality of criteria between List 1 and List 2 for rating the similarity of the low-meaningful stimulus pairs. The pairs of stimuli ranked highest are those with identical but transposed letter elements and the pairs differing only in their middle letters. The lowest-ranking pairs in each set are those where no elements are shared in the members of the pair.

Runquist and Joinson (in press) have scaled for similarity examples of all possible combinations of common elements among pairs of CCCs. The stimuli used in the present study represent 9 of the possible categories of common-element similarity for which Runquist and Joinson provide scale values; a rank correlation coefficient of 0.984 ($p < .01$) was found between the similarity values of the two studies. Thus the similarity of CCCs would appear to be largely predictable from the specification of the number and position of common letters shared by a pair of stimuli.

To summarize the findings in this section, evidence has been presented to show that subjects are able to rate the pair-wise similarity of verbal stimuli with reasonable internal consistency, although a significant decrease in consistency was found when subjects were asked to rank a set of stimuli whose meaningfulness differed from a homogeneous series that had been rated previously. Significant inter-set differences were found for internal consistency of response, although the differences seemed to be specific to a given set of stimuli rather than being governed by level of meaningfulness. Also, large individual differences were found in internal consistency of ranking.

Next, test-retest reliability coefficients calculated for the control material indicated that the reliability of the test was respectable,

TABLE XI

MEAN RANKING ASSIGNED TO STIMULUS PAIRS FROM LOW MEANINGFUL MATERIALS OF LIST 1 AND LIST 2

Pair	Mean Rank	Pair	Mean Rank
ZXJ - ZMJ	2.00	GQK - GXK	2.59
ZXJ - MZJ	4.43	GQK - XGK	5.56
ZXJ - DJZ	6.43	GQK - ZKG	7.10
ZXJ - WXQ	7.53	GQK - WGP	8.75
ZXJ - CGP	13.57	GQK - MHB	11.37
ZMJ - MZJ	2.42	GXK - XGK	2.15
ZMJ - DJZ	6.54	GXK - ZKG	4.82
ZMJ - WZQ	7.32	GXK - WGP	9.59
ZMJ - CGP	13.09	GXK - MHB	12.70
MZJ - DJZ	5.15	XGK - ZKG	3.45
MZJ - WZQ	7.07	XGK - WGP	6.95
MZJ - CGP	12.17	XGK - MHB	12.23
DJZ - WZQ	9.46	ZKG - WGP	8.90
DJZ - CGP	9.71	ZKG - MHB	12.89
WZQ - CGP	13.17	WGP - MHB	10.89

although the reliability for the high-meaningful material was significantly smaller than for the low-meaningful material. The hypothesis that a subject's ranking of the interstimulus distances could be changed by having him learn responses to the stimuli through a paired-associate procedure was not supported. Finally, it was shown that although significant communality is shown in subjects' ratings of all sets of stimuli, the ranking of the low-meaningful material seemed extremely stereotyped, while appreciable individual

differences were found in the ranking of the high-meaningful stimuli. The similarity of the low-meaningful stimuli seems to be predictable in terms of common elements.

Validity of Similarity Rankings

Having demonstrated that the reliability of the similarity ranking procedure is sufficient to warrant further study of the data, the next step is to examine the validity of the instrument. The content validity would seem obvious upon an examination of the instructions to the subjects and of the structure of the test, while determining predictive or concurrent validity is difficult due to the problem of specifying criteria for the similarity of the stimuli (except, perhaps, in the case of the low-meaningful material). The validation of the procedure, therefore, will concentrate on examining the construct validity by determining how the ranked interstimulus similarities relate to interference processes in paired-associate learning.

A decision had to be made as to whether all the available data should be used in an effort to stabilize estimates of experimental parameters with large sample sizes, or whether data selected for maximum reliability should be used. The latter alternative was chosen. It will be recalled that each subject did three ratings of experimental stimuli (ones which were used as paired-associate stimuli for that subject). One was done before the learning task, the second was a repetition of the first after the learning task, and the third was a rating of the remaining set of experimental stimuli that had not already been rated. It will also be remembered that the stimuli which were rated only once, after the learning task, elicited significantly lower internal consistency of response than all other ratings. For this reason, it was decided not to use these data in the val-

idation procedure. It was also found that the experimental material had shown some evidence of a decrease in reliability, compared to the control material. Although this difference was not large enough to be considered as adequate to reject the hypothesis that no difference exists in the population represented by the experimental samples, it is felt that this difference was sufficiently large to warrant these data's exclusion from consideration as a stable basis for further investigation. Accordingly, only the ratings of the experimental material done previous to the learning task were included in the tests which follow.

Relation Between Similarity Rankings and Confidence Ratings.

Two measures of stimulus generalization during learning were determined. The first was the confidence ratings given to the mismatched pairs on the recognition test. As was described earlier, each subject was presented with 16 incorrect pairings of stimuli and responses and required to give his extimate of the probability that the pairing was incorrect. Each stimulus-response mispairing can be considered to correspond to a particular stimulus pairing, this stimulus pairing consisting of the stimulus that had been presented in the mismatch and the stimulus associated with the response that had been presented in the mismatch. Eight of the stimulus-response mispairings from the recognition test thus correspond to stimulus pairs that had been ranked for similarity before the learning task, and so provide an opportunity to calculate a rank correlation coefficient expressing the relationship between rated similarity of a stimulus pair and an estimate of the strength of the generalization tendencies between two stimuli.

It had been hoped that an individual correlation coefficient could be calculated for each subject. However, this procedure did not provide much useful information, due to the high frequency of tied ranks.

It can be seen from Table XII that the subjects used only a few of the possible numbers between 0 and 100 in their responses on the recognition test; even after two learning trials almost half the ranks were tied. To obtain more stable estimates, means of the confidence ratings given to the first-ranked stimulus pairs, the second-ranked pairs, and so on, were calculated over the 16 groups of four subjects each who had been tested on the same material after the same number of trials of practice. Rank-order correlation coefficients were then calculated between the ranking of the stimulus pairs and the ranked mean confidence ratings.

Of the 16 groups, only three of the four groups who had been tested after four trials showed any significant correlation between similarity rankings and confidence ratings. The 8- and 16-trial groups had obviously learned most of the pairings correctly as they showed homogeneous and virtually perfect response, while the 2-trial group who had probably learned very little gave a wide, randomly-distributed spread of responses. Table XIII shows the 4-trial group's mean confidence ratings for the first-ranked stimulus-response pairing, the second-ranked pairing, and so on, with the ranking of the pairs determined by each subject's similarity ranking of the corresponding stimulus pairs. It also shows the value of the rank correlation coefficient for each group. Table XIII shows that the correlations are significant for three of four groups, and that the correlation for the combined data of all groups is quite substantial.

Relation between Similarity Rankings and Overt Response Confusion. The second measure of generalization was the frequency of response confusions shown on the first seven recall trials by 8-trial and 16-trial groups. This particular combination of number of trials and groups was used because it provided the largest total sample of response confusions where all subjects

TABLE XII

MEAN NUMBER OF CATEGORIES USED BY SUBJECTS
IN CONFIDENCE RATINGS

Trials	2	4	8	16
Mean	4.19	3.00	2.06	1.81
S. D.	0.94	1.06	1.25	1.02

TABLE XIII

MEAN CONFIDENCE RATINGS AND RANK CORRELATION COEFFICIENTS
FOR RANKED STIMULUS-RESPONSE MISPAIRINGS AFTER
FOUR LEARNING TRIALS

Meaningfulness	High		Low		Mean
Rank	1	2	1	2	
1	52.5	77.0	55.0	32.5	54.3
2	63.8	82.0	45.8	35.0	56.6
3	30.0	89.5	40.8	40.0	50.1
4	38.8	93.8	80.8	72.5	71.4
5	51.3	96.3	85.0	70.0	75.6
6	66.3	95.0	65.0	95.0	80.3
7	60.0	95.0	97.5	70.0	80.6
8	72.5	96.3	97.5	80.0	86.6
Mean	54.4	90.6	70.9	61.9	69.4
Rho	.50	.89**	.83*	.81*	.93**

* significant at the .05 level of probability

** significant at the .01 level of probability

had an equal chance to respond on all trials. By "response confusion" is meant the substituting of a response associated with another stimulus from within a given set of pairs for the correct response to a particular stimulus. If Stimulus A elicited the response that was paired with Stimulus B, or if the response for Stimulus A was given to Stimulus B, this error was attributed to a lack of discrimination between Stimulus A and Stimulus B. The number of response confusions that occurred to the highest-ranked interstimulus distance, to the second-ranked distance, and so on were then calculated. If two distances had been tied for a given ranking, then half an error was scored to each rank.

As only 166 errors of the type described above occurred on the trials considered, resulting in many tied ranks in the separate analysis of the four sets of stimuli, the data from the List 1 and List 2 subjects were combined to provide more stable estimates. The rankings of the mean ($N = 8$) frequency of response confusions attributed to each of the 15 rankings of interstimulus distances are shown in Table XIV for the high and low-meaningful stimuli, along with the associated rank correlation coefficients. It can be seen that the correlations for both sets of data are significant.

Effects of Similarity Rating Procedure upon Paired-Associate Learning

Having earlier examined the effects upon similarity ranking performance of prior practice at a paired-associate learning task, it would be of interest to study the converse situation. As each subject had rated for similarity half of the stimuli that occurred in the paired-associate list before starting practice, the opportunity is available to determine whether there are any differences in performance with these two sets of stimuli.

TABLE XIV

RANK CORRELATION OF FREQUENCY OF RESPONSE CONFUSION AND RANKING OF ASSOCIATED INTERSTIMULUS DISTANCE

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Rho
Meaningfulness																
High	5.0	4.0	9.0	5.0	6.0	7.0	9.0	3.0	3.0	4.0	3.0	2.0	3.0	1.0	1.0	0.786**
Low	12.0	8.0	10.0	5.5	9.5	10.0	2.0	6.0	16.0	6.0	3.0	5.0	4.0	1.5	2.5	0.703**
TOTAL	17.0	12.0	19.0	10.5	15.5	17.0	11.0	9.0	19.0	10.0	6.0	7.0	7.0	2.5	3.5	0.799**

** Significant at the .01 level of probability

As was described earlier, four groups of 16 subjects each practiced the paired-associate list for either 2, 4, 8 or 16 trials. However, as each subject had a recognition test after his last practice trial and a recall test after each practice trial before that, we actually have recall data from the four groups for 1, 3, 7 and 15 trials. Also, if we wished to examine performance after one practice trial, we would have N = 64, whereas for the data on Trials 8 - 15 we have N = 16. It is possible to examine the data over 15 trials with the realization that N decreased as trials increased, but it would be difficult to perform a valid analysis of variance on data of this type. Various combinations of groups will be selected for analysis, therefore, so that all subjects have equal opportunity to contribute to the data under all conditions. It should be realized, of course, that this decision necessitates that a compromise must be reached on any given analysis between the number of subjects included and the number of trials over which the analysis extends, according to the hypothesis being considered.

The first analysis will include data from all 15 trials from 16 subjects. Table XV shows the summary of an analysis of variance on the mean number of correct responses to each set of stimuli on each recall test as a function of list, meaningfulness, trials, and whether or not the stimuli had been rated for similarity previously (familiarization).

Table XV shows that only the effects of meaningfulness, trials, and the interaction between meaningfulness and trials are significant. Figure 4 gives the data for these conditions graphically. It can be seen that the high-meaningful material elicits more correct responses than the low-meaningful material, performance improves as a function of trials, and the performance for both levels of meaningfulness approaches the asymptote

of perfect performance near the end of practice, all findings that are highly predictable from previous data.

It can be seen that the F-ratio for the factor of familiarization is substantial, although not significant. As was outlined earlier in this paper, it is expected that the effects of familiarization are greatest in the early stages of practice. It is thus possible that familiarization has a significant effect on the data of the earlier trials, but that this effect is obscured when the data over all trials are summed. This hypothesis was tested by applying the same analysis of variance design to the data of Trials 1 - 7 only, which also permits a more stable estimate of the effects of the experimental variables by increasing the number of subjects to N = 32.

Table XVI shows a summary of this analysis. It can be seen that meaningfulness and trials effects are still significant. It can also be seen that familiarization has had a significant effect. From Figure 5 it can be seen that the familiarized set of stimuli elicit fewer correct responses than do the non-familiarized stimuli. It would seem reasonable to conclude from the data of Trials 1 - 7 that familiarization has had a depressing effect on the association of responses with stimuli in the early portion of the learning task. Whether the absence of this difference in the data of Trials 1 - 15 is due to an obscuring of the effect by the additional trials or to the lesser sensitivity of this analysis due to the smaller sample cannot be determined.

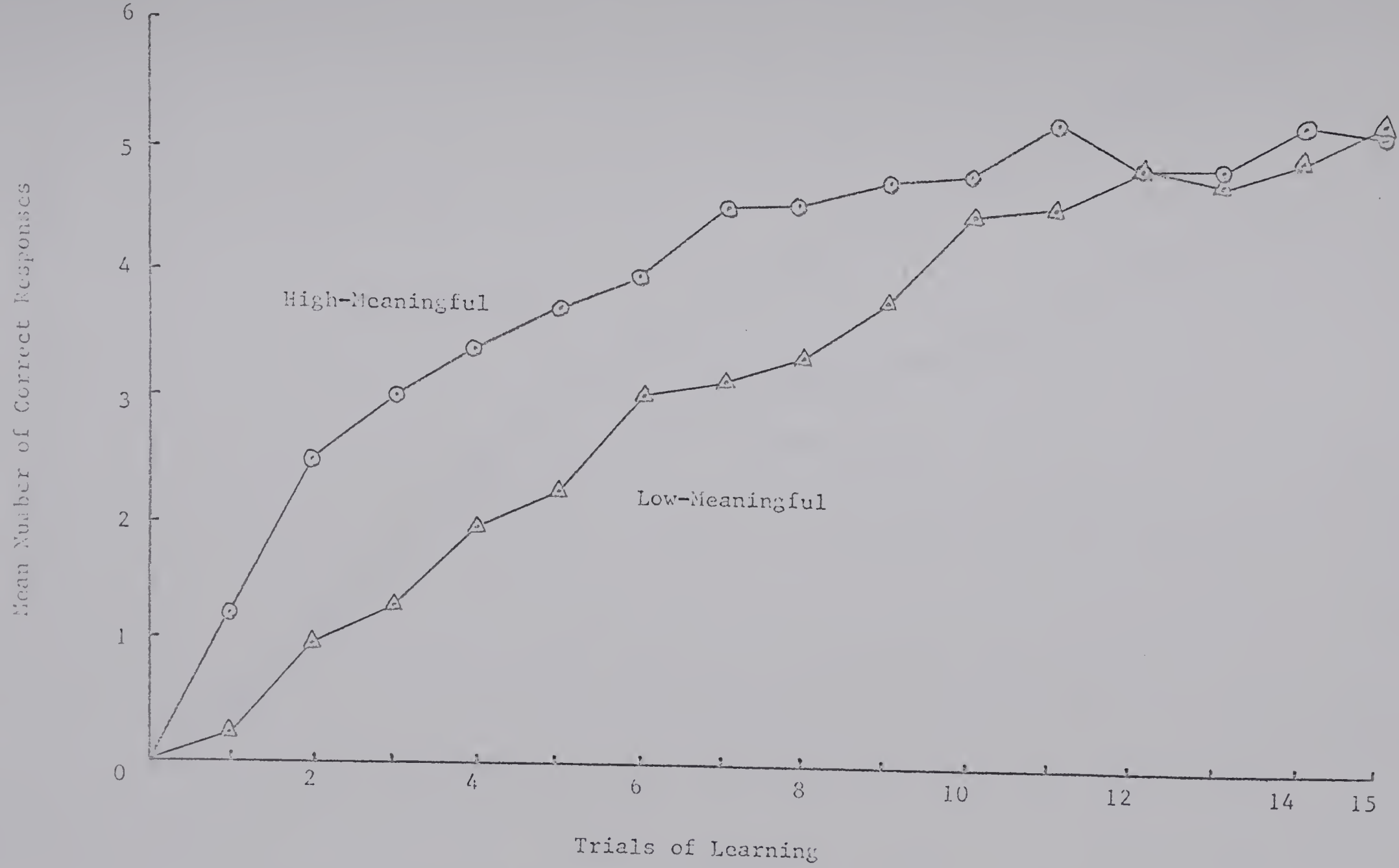


Figure 4. Mean number of responses correctly recalled over fifteen trials as a function of meaningfulness and trials

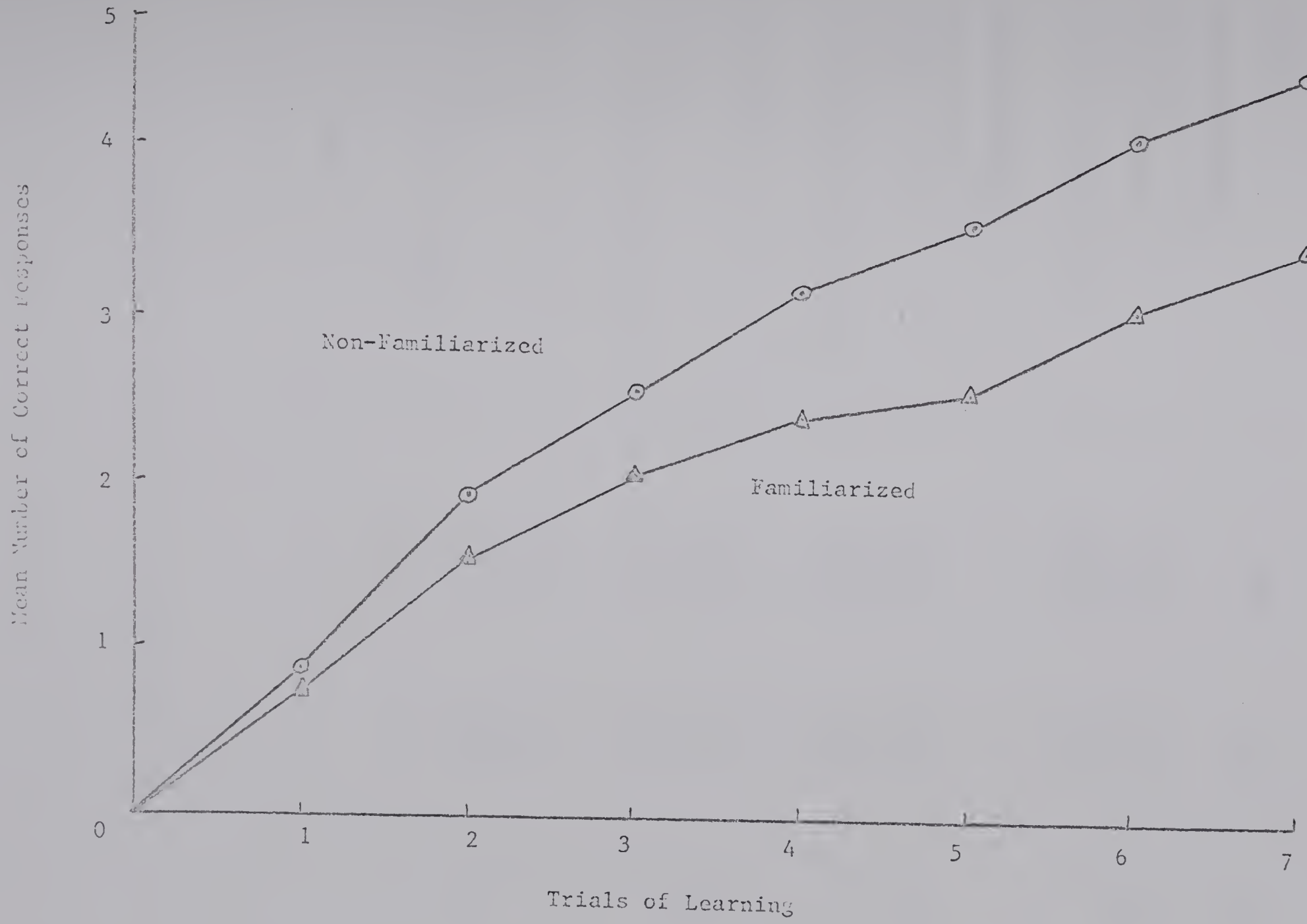


Figure 5. Mean number of responses correctly recalled over seven trials as a function of experimental condition and trials.

TABLE XV

SUMMARY OF ANALYSIS OF VARIANCE OF MEAN NUMBER OF CORRECT RESPONSES ON RECALL TESTS OVER FIFTEEN LEARNING TRIALS

Source	df	SS	MS	F
Between subjects				
List (L)	1	75.208	75.208	2.32
Familiarization (F)	1	93.633	93.633	2.89
L X F	1	9.633	9.633	< 1
Subjects within groups	12	388.383	32.365	-
Within subjects				
Meaningfulness (M)	1	83.333	83.333	15.72**
L X M	1	9.633	9.633	1.82
F X M	1	3.675	3.675	< 1
L X F X M	1	3.008	3.008	< 1
M X Subjects within groups	12	63.617	5.301	-
Trials (T)	14	834.992	59.642	54.28**
L X T	14	20.792	1.485	1.35
F X T	14	22.992	1.642	1.49
L X F X T	14	2.742	0.196	< 1
T X Subjects within groups	168	184.617	1.099	-
M X T	14	42.792	3.057	3.77**
L X M X T	14	8.242	0.589	< 1
F X M X T	14	12.825	0.916	1.13
L X F X M X T	14	13.492	0.964	1.19
M X T X Subjects within groups	168	136.383	0.812	-
Total	479	-	-	-

** Significant at the .01 level of probability

TABLE XVI

SUMMARY OF ANALYSIS OF VARIANCE OF MEAN NUMBER OF CORRECT RESPONSES ON RECALL TESTS OVER SEVEN LEARNING TRIALS

Source	df	SS	MS	F
Between subjects				
List (L)	1	29.009	29.009	2.379
Familiarization (F)	1	52.938	52.938	4.342*
L X F	1	6.036	6.036	< 1
Subjects within groups	28	341.411	12.193	-
Within subjects				
Meaningfulness (M)	1	198.223	198.223	45.653**
L X M	1	9.143	9.143	2.106
F X M	1	0.893	0.893	< 1
L X F X M	1	0.437	0.437	< 1
M X Subjects within groups	28	121.589	4.342	-
Trials (T)	6	447.085	74.514	76.898**
L X T	6	6.022	1.004	1.036
F X T	6	11.906	1.984	2.047
L X F X T	6	1.433	0.239	< 1
T X Subjects within groups	168	162.839	0.969	-
M X T	6	5.246	0.874	1.118
L X M X T	6	9.201	1.533	1.960
F X M X T	6	6.513	1.086	1.389
L X F X M X T	6	6.344	1.057	1.352
M X T X Subjects within groups	168	131.411	0.782	-
Total	447	-	-	-

* Significant at the .05 level of probability

** Significant at the .01 level of probability

CHAPTER IV

Discussion

In the following section the major findings of the study will be reviewed. This will be followed by a discussion of some questions arising from the results.

Review of Results

The procedure of this study had subjects rank for similarity 60 out of 105 possible pairs of distances between six stimuli. Through the use of appropriate assumptions it was found that a ranking could be derived for over 99% of the total 105 pairs of distances from the information contained in the 60 rated pairs. There was enough redundancy in the experimental data to provide a partial check on one of these assumptions; that the common ordering of the 15 distances that made up the 105 pairs was transitive. Fewer than 5% of the observed rankings conflicted with this assumption. The majority of subjects apparently show high internal consistency in their similarity ratings.

The effect upon subjects' consistency of rating was studied as a function of several variables. Reliable differences in consistency of response over several ratings were demonstrated for individual subjects. Mean consistency of response dropped when the meaningfulness of the material was changed after the subject had rated several stimulus sets at a given level of meaningfulness. Although the two sets of stimulus material had been selected so as to be as equivalent as possible, sample-specific differences in consistency of rating were found. These differences were not related to the meaningfulness *per se* of the material, however. Nor were any

differences found as a result of previously using the stimuli in a paired-associates learning task.

The test - retest reliability of the similarity rankings was found to be moderately high; the reliability of the low-meaningful material was greater than that of the high-meaningful material. No significant change in reliability was found as a result of using the stimuli in a paired-associates learning task between ratings.

It was found that there was a high degree of agreement between subjects on the similarity ratings of the low-meaningful material. Furthermore, a correlation of virtually unity was found between the mean similarity ratings of the low-meaningful stimuli in this experiment and those of another similarity rating experiment, even though different material and different rating procedures were used. Apparently common standards of similarity for low-meaningful material prevail in the population sampled in the present experiment. A low level of concordance of rating was found between subjects for the high-meaningful material, however, even when a correction was made for the reliability of the rating procedure. This would seem to indicate that subjects' standards for the similarity of the meaningful material studied in this experiment tended toward the idiosyncratic.

No significant relationship was found between stimulus similarity and confusion errors on the recognition test after 2, 8, and 16 trials of training. After four training trials, however, three out of four sets of stimuli (one high-meaningful, two low-meaningful) and the pooled data of the four sets showed a reliable relationship between rated similarity and confusion errors. In addition, the pooled data for the high-meaningful and the low-meaningful stimuli showed a significant relationship between rated

similarity and overt intra-list intrusions over the first seven trials. It was also found that subjects associated fewer correct responses on Trials 1 - 7 to stimuli that were rated before practice on the learning task than to stimuli that were rated after learning.

Evaluation of Similarity Ranking Test

It would seem that the similarity-rating procedure studied in this experiment has some merit as a tool in future research. The test was shown to have reasonable internal consistency and test-retest reliability over short periods for material of high and low meaningfulness, and to be able to predict well the occurrence of both intra-list intrusions and recognition errors in paired-associate learning.

It is even more encouraging to note the potentialities for improvement that seem possible in the test. As was stated earlier, a compromise between quality and quantity was struck in the planning of this experiment--in order to determine the effects of a number of variables without demanding too much in terms of time from the subjects, a relatively short version of the similarity-rating procedure was used. A minimum of over-determination of the data was planned, with 60 out of a possible 105 paired comparisons being made, each once only.

If it were wished, however, to get a more reliable inter-stimulus distance ranking for a few stimuli in the same amount of time in preference to obtaining data on a larger number of stimuli, there would seem to be no reason why this could not be done. The procedure used in this study might be repeated a number of times to obtain a more stable estimate of the rankings. What would probably be better would be to use another of the cartwheel data collection methods discussed by Coombs (1964) that provide more redundant information, and thus also more thoroughly check the internal

consistency of the data. The present study used what Coombs terms a Case 1 cartwheel design, where the basic test involves all three possible comparisons of pairs among three stimuli. The use of a Case 3 cartwheel design (where a given "hub" stimulus is paired with three "rim" stimuli on the basic test and the three combinations of pairs are ranked) in a ranking test would give three independent rankings of each pair of stimulus pairs (each ranking in a different context), and presumably take three times as long for the subject to do.

Another aspect of the evaluation of this test that would seem fertile ground for improvement is the recognition test procedure. It will be recalled that the similarity ranking data were validated in part by correlation with the subjects' ratings of their certainty that a pair had been changed. One of the disappointments of this experiment was the failure of the majority of the subjects to make use of more response categories on the recognition test. As a result, data from groups of four subjects had to be pooled to provide reasonable variation in this estimate of generalization.

One possible explanation for this finding is that the subjects showed response bias due to their inexperience with making the type of probability judgment required in this experiment. This possible response bias might be eliminated by giving subjects practice in probability estimation prior to the experiment. Alternatively, a different form of response might be used by subjects to indicate judgments, such placing a mark on a line (cf. Garskof and Houston, 1963). Assuming that a change in procedure would produce greater variation in judgments, it would remain to be determined, of course, whether fineness of judgment increased as well.

The Determinants of Similarity

Besides the validation of the similarity rating procedure, the present study provides some other interesting information, especially on the determinants of similarity in material of high and low meaningfulness. Runquist & Joinson's (in press) finding that rated similarity of low-meaningful material is determined by the sharing and position of common elements was confirmed. Although this study did not use as large or representative a sample of stimuli as did Runquist and Joinson, the correlation between the mean similarity ratings of the stimulus-types that did occur in both studies was virtually unity. The similarity of the low-meaningful material used in this study seems to be determined by highly stereotyped standards. These standards seem to be predictable in terms of a common-elements theory of similarity, where the elements are letters.

The same cannot be said about the ratings of the high-meaningful material, however. This study has presented evidence that similarity rankings of high-meaningful verbal stimuli, when compared to low-meaningful stimuli, show (a) less inter-subject consistency of response, i.e., agreement among subjects about a stereotyped ranking, (b) poorer short-term test-retest reliability within subjects, (c) no difference in internal consistency of ranking. This last point is quite important, because it apparently eliminates a simple explanation of the first two effects, that similarity of meaningful words, as measured by the present technique, is essentially unstable. The fact that subjects are capable of showing short-term consistency of response when ranking high-meaningful stimuli equal to that shown while ranking low-meaningful material would seem to require a more complex explanation. The results would seem to support the hypothesis presented earlier in this paper that subjects differ in their criteria in

judging similarity of meaningful words because the characteristics which determine meaningful similarity are learned, and hence these criteria vary between subjects with different experience. They also seem to indicate that no more change in judgment criteria is found during a given ranking test for high-meaningful material than for low-meaningful material, but that in approximately half an hour subjects will change to a significantly different set of criteria, which show no change in internal consistency. However, although this interpretation of the data would seem to be the simplest available, several points should be examined carefully.

First, this interpretation involves, in part, accepting the null hypothesis, a dubious procedure. It means that we consider that the observed lack of significant difference in internal consistency between the rankings of the samples of high and low-meaningful material is representative of the populations from which that these samples were drawn. However, it will be recalled that the low-meaningful material showed slightly, although not significantly, greater internal consistency. Second, it will also be recalled that significant differences were found in internal consistency within the four sets of similarity rankings, although these seemed to be specific to the sets and not related to level of meaningfulness or list alone.

It would seem, then, that we can conclude that the evidence favors the hypotheses that the rankings of the high- and low-meaningful materials in this experiment differ in both between-subjects consistency and test-retest consistency over short intervals. However, no compelling evidence could be found concerning differences in internal consistency of the ranking of a single set of stimuli.

The conclusion that high inter-subject variability is typical of

similarity rankings for high-meaningful material might also be studied more carefully as well, but for different reasons. There is no reason to believe that serious sampling error occurred when the high-meaningful material was selected. However, it must be remembered that only one limited class of meaningful words - adjectives - was sampled from. It is entirely possible that verbal units with a more concrete denotative function (e.g., concrete nouns) might be ranked with much greater inter-subject consistency. This point becomes all the more worthy of consideration when it is recalled that only high-frequency adjectives were used. It would seem quite plausible that words that are widely used in everyday experience might be so selected because of their flexibility and versatility of semantic function, and that adjectives with a narrower range of usage might be represented by less variable points in semantic space.

Familiarization of Stimuli

The results of this study indicated that stimuli, both of high and low meaningfulness, which had been rated for similarity previous to the learning task had fewer responses learned to them than did the non-rated stimuli in the initial stages of practice on a paired-associate list. The results of this study support others that have presented evidence contrary to the hypothesis that experience with stimulus items prior to their use in a paired-associate learning task facilitates the learning of responses to these stimuli.

There are a number of possible explanations for the observed disconfirmation of this hypothesis. Previous experience with stimuli is presumed to facilitate learning because it provides the subject with an opportunity to improve his discrimination of the stiruli prior to actual practice with the list. Two schools of thought exist as to the nature of

this proposed prior discrimination. One maintains that distinctive responses are learned to the stimuli, thus increasing their differentiation by the creation of new, more easily-discriminated stimulus complexes. The other postulates that prior experience with the stimuli enables the subject to learn to attend to their distinguishing features and thus eliminates some of the confusion between stimuli that previously existed. According to the first theory any operation which attaches distinctive responses to the stimuli will facilitate subsequent paired-associate learning. The second theory maintains that associating distinctive responses to the stimuli is not necessary for subsequent facilitation; any operation that encourages the subject to attend to the distinctive features of each stimulus would achieve this end.

Three hypotheses, related to the above theories, might explain why inhibition, rather than facilitation, was found in this study. The first hypothesis is that associations to the stimuli acquired through similarity rating produced more direct interference with the responses to be learned on the paired-associate task than they produced facilitation through acquired distinctiveness. In other words, the responses learned to the stimuli in the acquired-distinctiveness training, although they may have produced some facilitation by making the stimuli more discriminable, produced a net inhibition effect by intruding during paired-associate practice and blocking the acquisition of the correct responses. Jung (1967) has suggested that familiarization techniques cause incidental associations between the items being familiarized. These incidental associations in turn cause interference with the learning of responses in the paired-associates task. It is quite possible that this is an explanation for the inhibition of learning found in this experiment.

A second hypothesis also has some plausibility when the procedure of the present study is considered. As Goss and Nodine (1965) have pointed out, one of the operations that has been used to produce acquired distinctiveness - simple presentation of the stimulus on the assumption that eliciting a "recognition response" will cause increased integration of this response, and increase the stimulus' discriminability - is the same as the operation used to produce "semantic satiation" (Lambert & Jakobovits, 1960). Semantic satiation presumably involves a loss of meaning for the stimulus, which presumably would hinder the association of a response with it.

It is difficult to conceptualize the mechanism of semantic satiation producing the results observed in this experiment when it is recalled that no difference in inhibitory effects were observed between the high-meaningful stimuli, which presumably have a good deal of meaning to lose, and the low-meaningful stimuli, which by definition are virtually meaningless. However, it is possible that habituation involving a novelty rather than a semantic factor might be playing a significant part here. It will be recalled that half of the stimuli on the paired-associate list had been familiarized (by similarity ranking), while the other half had not. It is possible that the "new" stimuli might have elicited novelty reactions that in some way facilitated the association of a response, while the familiarized stimuli would not be perceived as vividly and would not be attended to as intensely as the "new" stimuli. However, when it is remembered that the "new" and "old" stimuli were always of opposite extremes of meaningfulness for each subject, it would seem likely that differential effects of familiarization would be observed between the low-meaningful stimuli, which should have a high initial novelty value and hence more to lose, and the high-meaningful stimuli, whose level of familiarization would presumably be near

asymptote. This difference was not observed in the data of this experiment. It may be possible that the test of this hypothesis was not sensitive enough to reveal this difference, but this particular explanation of the experimental data will remain uncertain until this anomaly can be adequately resolved by supporting evidence.

A third possible explanation for the effects of the familiarization procedure concerns the nature of the similarity ranking task. It was mentioned earlier that one theory postulates that discrimination between stimuli improves with experience with the stimuli because the subject attends more to the distinctive features of each stimulus as practice increases, and learns to ignore or not respond to features that do not distinguish different stimuli. This attention to distinctive features is obviously beneficial in paired-associate learning where a different response must be associated to each stimulus. But it is quite likely that the similarity ranking task, far from encouraging subjects to search for and attend to distinctive features of the stimuli, actually encouraged the opposite: searching for and attending to <u>similar</u> features of the stimuli, while ignoring the unique and hence distinctive features. This, of course, is what the instructions required the subjects to do, and it is possible that sufficient practice at searching for similarities would produce a form of acquired equivalence of stimuli through an unlearning of mediating responses to the unique features of each stimulus.

This last-mentioned hypothesis suggests an experimental test. Subjects could be given a modified form of the instructions used in the interstimulus-distance ranking portion of this experiment by changing the word "similar" to "different" and the word "similarity" to "difference". This would change the procedure to a difference-rating rather than a

similarity-rating one. By comparing the learning scores of subjects who underwent this modified procedure to the scores of control subjects and of subjects who underwent a replication of the procedure of the present study, data relating to two questions could be obtained:

(1) Do subjects who participate in a difference-ranking procedure (acquired distinctiveness training) subsequently learn responses to the rated stimuli more quickly than controls, and do subjects who attempt a similarity-ranking task (acquired equivalence training) learn responses more poorly?

(2) Is an interstimulus-distance ranking scale obtained through similarity ratings different from one derived from difference ratings?

Reduction of Intra-list Interference

It has been demonstrated that the average person is capable of remembering perfectly a single pair of words, or even several pairs, after he has been presented with them only once (Miller, 1956). However, when over a certain number of pairs are presented once and the memory of the pairings subsequently tested it is typically found, as in this experiment, that recall is less than perfect. Presumably attempting to learn several things in close temporal proximity produces mutual interference among the separate items.

Gibson (1940) has proposed that an important aspect of this interference is caused by generalization between the stimulus items in the paired-associates list. According to her hypothesis, before the subject's first attempt to learn a paired-associates list the associative bonds between stimuli and responses should be negligible if he has not experienced the pairings before. After he has been exposed to the pairs for a few trials, the strength of the association between each stimulus and its correct response will increase, resulting in a growth of habit strength of the

correct responses. However, if the stimuli are similar to each other the subject will confuse them, causing him to associate wrong (or generalized) responses to a given stimulus as well. In this way the habit strength of both the correct and the incorrect responses will increase during the initial trials of learning, resulting in interference with the correct responses. But as the frequency of elicitation of the correct responses increases, and with it the frequency of the generalized ones, differentiation between the stimuli will also increase. This is because the correct responses are reinforced when they are produced, increasing their habit strength, while the incorrect responses are not reinforced, resulting in their eventual extinction. The net effect, according to Gibson, will be that interference due to stimulus generalization will increase in the early stages of paired-associates learning with the growing habit strength of the correct responses. It will reach a peak at some intermediate stage of practice and then subsequently decrease as differentiation increases.

Gibson (1942) proposed that frequency of overt intra-list response intrusions be used as an index of stimulus generalization in paired-associates learning. As the typical paired-associates learning experiment shows a rise, then a fall, in frequency of occurrence of overt response intrusion errors as a function of learning trials (Murdock, 1958), until recently it has been considered that Gibson's hypothesis concerning the rise and fall of stimulus generalization has, in general, been supported by data. However, Murdock (1958) has pointed out that the process of response production, correct or incorrect, is usually imperfect at the beginning of the typical paired-associates experiment and increases with practice. It is generally considered (Underwood and Schulz, 1960) that integration or differentiation of the response items could interfere with the formation of stimulus-response

bonds in the early portions of the typical paired-associates learning procedure. A valid test of Gibson's hypothesis should, according to Murdock (1958), employ some means of avoiding the possible confounding effect of response acquisition in the early stage of paired-associates learning.

Muir (1963) used a recognition procedure similar to the one of the present study in an attempt to achieve this end. After varying numbers of practice trials on a paired-associates list subjects were presented with correct and incorrect pairings and asked if the pairings were the same as or different to the ones on the practice list. The subjects were told to respond only if they were "reasonably sure" that they could or could not detect a change, and to respond accordingly if they were uncertain. It was found that misrecognitions of incorrect pairings did not increase and then decrease as would be expected from Gibson's (1940) theory, but instead were at a maximum after the first practice trial and decreased monotonically thereafter.

The results of the present experiment, however, could be interpreted as support for Gibson's (1940) hypothesis. Although many misrecognitions of incorrect pairings occurred after the first recognition test (i.e., after the second practice trial), no reliable connection between rated stimulus similarity and recognition errors could be demonstrated for these data. This would seem to indicate that stimulus generalization could not be used as an explanation for errors at this point (Trial 2) in learning. But a reliable relationship was shown between stimulus similarity and recognition errors for the data of the second recognition test, after Trial 4. For the two subsequent tests after Trial 8 and Trial 16 virtually no recognition errors occurred.

It could be concluded, then, that although both the present study and Muir (1963) show that recognition errors for mispairings in paired-associates learning decrease monotonically as a function of practice, the data of this study seem to indicate that misrecognitions which can be reliably attributed to stimulus generalization are at a maximum at some intermediate stage of practice. However, no explanation seems apparent for the large number of misrecognitions that occurred on Trial 2 of the present study, and for the fact that Muir (1963) could not reliably demonstrate that stimulus generalization had been a factor in causing recognition errors. It seems apparent that variables other than stimulus generalization have an effect on recognition errors in the early stages of paired-associates learning, and it is possible that the effects of stimulus generalization are obscured and hence not detectable after Trial 2 of the present study. The nature of these other variables is not apparent from the data presented here.

References

Abbott, D.W., Effects of meaningfulness of structurally similar CVCS on stimulus generalization of eyelid closure. J. exp. Psychol., 1966, 71, 511-515

Abbott, D.W., & Price, L.E., Stimulus generalization of the conditioned eyelid response to structurally similar nonsense syllables. J. exp. Psychol., 1964, 68, 368-371.

Amster, Harriett, Semantic satiation and generation: Learning? Adaptation? Psychol. Bull., 1964, 62, 273-286.

Attneave, F., Dimensions of similarity. Amer. J. Psychol., 1950, 63, 516-556.

Attneave, F. Ability to verbalize similarities among concepts and among visual forms. Amer. Psychologist, 1951, 6, 270 (Abstract).

Baddely, A.D., Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. Quart. J. exp. Psychol., 1966, 18, 362-365.

Bastian, J., Associative factors in verbal transfer. J. exp. Psychol., 1961, 62, 70-79.

Bousfield, W.A., The occurrence of clustering in the recall of randomly arranged associates. J. gen. Psychol., 1953, 49, 229-240.

Bousfield, W.A., Whitmarsh, C.A., & Berkowitz, H., Partial response identities in associative clustering. J. gen. Psychol., 1960, 63, 233-238.

Cofer, C.N., Associative commonality and rated similarity of certain words from Haagen's list. Psychol. Rep., 1957, 3, 603-606.

Conrad, R., An association between memory errors and errors due to acoustic masking of speech. Nature, 1962, 193, 1314-1315.

Conrad, R., Acoustic confusions in immediate memory. Brit. J. Psychol., 1964, 55, 75-84.

Coombs, C.H., A theory of data. New York: Wiley, 1964.

Dallett, K.M., Effects of within-list and between-list acoustic similarity on the learning and retention of paired associates. J. exp. Psychol., 1966, 72, 667-677.

Dember, W.H., The relation of decision-time to stimulus similarity. J. exp. Psychol., 1957, 53, 68-72.

Dicken, C.F., Connotative meaning as a determinant of stimulus generalization. Psychol. Monogr., 1961, 75, (1, Whole No. 505).

Feldman, S.M., & Underwood, B.J., Stimulus recall following paired-associate learning. J. exp. Psychol., 1957, 53, 11-15.

Flavell, J.H., Meaning and meaning similarity: I. A theoretical reassessment. J. gen. Psychol., 1961a, 64, 307-319.

Flavell, J.H., Meaning and meaning similarity: II. The semantic differential and co-occurrence as predicters of judged similarity in meaning. J. gen. Psychol., 1961b, 64, 321-335.

Flavell, J.H., & Johnson, Ann B., Meaning and meaning similarity: III. Latency and number of similarities as predicters of judged similarity in meaning. J. gen. Psychol., 1961, 64, 337-348.

Gannon, D.R., & Noble, C.E., Familiarization (n) as a stimulus factor in paired-associate verbal learning. J. exp. Psychol., 1961, 62, 14-23.

Garskof, B.E., & Houston, J.P., Measurement of verbal relatedness: An idiographic approach. Psychol. Rev., 1963, 70, 277-288

Gibson, Eleanor J., A systematic application of the concepts of generalization and differentiation to verbal learning. Psychol. Rev., 1940, 47, 196-229.

Gibson, Eleanor J., Intra-list generalization as a factor in verbal learning. J. exp. Psychol., 1942, 30, 185-200.

Goss, A.E., & Nodine, C.F., Paired-associates learning: The role of meaningfulness, similarity and familiarization. New York: Academic Press, 1965.

Haagen, C.H., Synonymity, vividness, familiarity, and association-value ratings for 400 pairs of common adjectives. J. Psychol., 1949, 27, 453-463.

Helm, C.E., & Tucker, L.R., Individual differences in the structure of color-perception. Amer. J. Psychol., 1962, 25, 437-444.

Higa, M., Interference effects of intralist word relationships in verbal learning. J. verb. Learn. verb. Behav., 1963, 2, 170-175.

James, W., Principles of psychology. New York: Holt, 1890.

Jenkins, J.J., Degree of polarization and scores on the principal factors for concepts in the semantic atlas study. Amer. J. Psychol., 1960, 73, 274-279.

Jung, J., Transfer analysis of familiarization effects. Psychol. Rev., 1967, 74, 523-529.

Kausler, D.H., (Ed.) Readings in verbal learning: Contemporary theory and research. New York: Wiley, 1966.

Kurcz, Ida, Semantic and phonetographic generalization of a voluntary response. J. verb. Learn. verb. Behav., 1964, 3, 261-268.

Lambert, W.E., & Jakobovits, L.E., Verbal satiation and changes in the intensity of meaning. J. exp. Psychol., 1960, 60, 376-383.

McGeoch, J.A. & McDonald, W.T., Meaningful relation and retroactive inhibition. Amer. J. Psychol., 1931, 43, 579-588.

Marshall, G.R., & Cofer, C.N., Associative indices as measures of word relatedness: A summary and comparison of ten methods. J. verb. Learn. verb. Behav., 1963, 1, 408-421.

Miller, G.A., The magical number seven plus or minus one: Some limits on our capacity for processing information. Psychol. Rev., 1956, 63, 81-97.

Morgan, R.L., & Underwood, B.J., Proactive inhibition as a function of response similarity. J. exp. Psychol., 1950, 40, 592-603.

Muir, W.R., Stimulus generalization as a function of training in verbal learning, as measured by a recognition test. Unpublished M.A. thesis, University of Manitoba, 1963.

Murdock, B.B., Jr. Intra-list generalization in paired-associate learning. Psychol. Rev., 1958, 65, 306-314.

Noble, C.E., An analysis of meaning. Psychol. Rev., 1952, 59, 421-430.

Osgood, C.E., The similarity paradox in human learning: A resolution. Psychol. Rev., 1947, 56, 132-143.

Osgood, C.E., The nature and measurement of meaning. Psychol. Bull., 1952, 49, 197-237.

Osgood, C.E., Method and theory in experimental psychology. New York: Oxford, 1953.

Osgood, C.E., Studies in the generality of affective meaning systems. Amer. Psychologist, 1962, 17, 10-28.

Osgood, C.E., & Suci, C.J., A measure of relation determined by both mean difference and profile information. Psychol. Bull., 1952, 49, 251-262.

Osgood, C.E., & Suci, C.J., & Tannenbaum, P.H., The measurement of meaning. Urbana, Ill; University of Illinois Press, 1957.

Postman, L., The generalization gradient in recognition memory. J. exp. Psychol., 1951, 42, 231-235.

Razran, G.H.S., A quantitative study of meaning by a conditioned salivary technique (semantic conditioning). Science, 1939, 90, 89-90.

Razran, G., Semantic and phonetographic generalization of salivary conditioning to verbal stimuli. J. exp. Psychol., 1949, 39, 642-652.

Richardson, J., The relationship of stimulus similarity and number of responses. J. exp. Psychol., 1958, 56, 478-484.

Riess, B.F., Semantic conditioning involving the galvanic skin reflex. J. exp. Psychol., 1940, 26, 238-240.

Rowan, T.C., Some developments in multidimensional scaling applied to semantic relationships. Unpublished doctoral dissertation, University of Illinois, 1954.

Runquist, W.N., & Joinson, Peggy, Rated similarity of trigrams. J. verb. Learn. verb. Behav., 1968, in press.

Ryan, J.J., Comparison of verbal response transfer mediated by meaningfully similar and associated stimuli. J. exp. Psychol., 1960, 60, 408-415.

Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Slamecka, N.J., Transfer with mixed and unmixed lists as a function of semantic relations. J. exp. Psychol., 1967, 73, 405-410.

Staats, Carolyn K., & Staats, A.W., Meaning established by classical conditioning. J. exp. Psychol., 1957, 54, 74-80.

Thorndike, E.L., & Lorge, I., The teacher's word book of 30,000 words. New York: Columbia University Press, 1944.

Tighe, Louise S., & Tighe, T.J., Discrimination learning: two views in historical perspective. Psychol. Bull, 1966, 66, 353-370.

Torgerson, W.S., Theory and methods of scaling. New York: Wiley, 1958.

Underwood, B.J., Associative transfer in verbal learning as a function of response similarity and degree of first list learning. J. exp. Psychol., 1951, 42, 44-53.

Underwood, B.J., Studies of distributed practice: IX. Learning and retention of paired adjectives as a function of intralist similarity. J. exp. Psychol., 1953, 45. 143-149.

Underwood, B.J., & Schulz, R.W., Meaningfulness and verbal learning. Chicago: Lippincott, 1960.

Vastenhouw, J., Relationships between meanings. The Hague: Mouton, 1962.

Wallach, M.A., On psychological similarity. Psychol. Rev., 1958, 65, 103-116.

Wickelgren, W.A., Acoustic similarity and retroactive interference in short-term memory. J. verb. Learn. verb. Behav., 1965, 4, 53-61.

Wickelgren, W.A., Short-term recognition memory for single letters and phonemic similarity of retroactive interference. Quart. J. exp. Psychol., 1966, 18, 55-62.

Wimer, Cynthia, An analysis of semantic stimulus factors in paired-associate learning. J. verb. Learn. verb. Behav., 1963, 1, 397-407.

Winer, B.J., Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Witmer, L.R., The association value of three-place consonant syllables. J. genet. Psychol., 1935, 47, 337-360.

Appendix A

Sample Stimulus Similarity Rating Form

Appendix B

Instructions for Rating Similarity of Stimuli

On the following pages are a number of sets of three words. Each set will be arranged in a triangle like the example below. You are to decide:

(a) which two of the three words are most similar to each other,

(b) which two of the three words are least similar to each other.

Indicate your choice by putting a plus sign (+) in the circle between the two words that you think are most similar, and a minus sign (-) in the circle between the two words you consider to be least similar.

It is up to you to decide what property of the words you will use to judge their similarity by. However, don't spend too much time in making a decision-- your first impression is the best basis of judgment. Also, consider only one set of words at a time -- don't look back at or change your previous ratings.

In the example below, the three words can be considered as three pairs -- (a) Word 1 - Word 2, (b) Word 2 - Word 3, and (c) Word 1 - Word 3. If you think that Word 2 and Word 3 are the most similar of the three pairs, mark a plus between them as shown below. Similarly, put a minus between Word 1 and Word 3 if you consider them as the least similar. By elimination, of course, Word 1 and Word 2 would be of intermediate similarity.

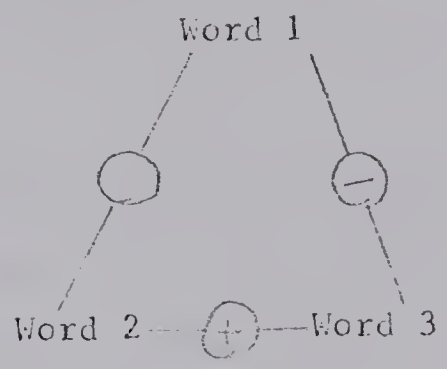
Remember,

(1) It is up to you to decide in what way the words are similar.

(2) Work quickly but carefully.

(3) Consider only one set of words at a time when making a rating -- don't

look back at or change your previous judgments.



Appendix C

Instructions for Learning Paired-Associate List

Initial Instructions

I want you to learn a list of word-pairs that I'm going to show you. You're to try to learn the pairings of the words. I'll present each pair of words for two seconds on the screens in front of you. During this presentation period I want you to study the pairings. Try to learn as many of the pairings as possible with the objective of learning them all. After this study trial the screen will be blank for a few seconds, and then I'll test you to see how many of the pairings you can remember. Following this test, the pairings will again be presented for two seconds each for you to study, followed by another test. So the procedure runs, study-test-study-test, and so on.

I'm going to test your memory for the pairings in two ways. In the first type of test, called a recall test, I'll simply present the first word in a pairing, and you'll have four seconds in which to try to remember and call out the second word of the pair. So every time you see only the first word of a pair presented, try to remember the second word and call it out. Unless I tell you otherwise, a recall test will follow every study trial, so be prepared to respond when the screen goes blank, as you'll only have four seconds for each response.

In the second type of test, called a recognition test, I'm going to scramble some of the pairs and see if you can recognize which ones have been changed. In other words, I'm going to take the words in some of the pairs that I've shown you earlier and switch the pairings around. I'll then show you, one at a time, some of these new pairings and some of the

old pairings. As I show you each test pair, I want you to give me your estimate of the percent probability that the pairing has been changed. You should give me this estimate as a number between zero and one hundred, a large number indicating that you think that it's probable that the pairing has been changed, and a small number showing that it's probable that the pairing has not been changed. This procedure is the same as one that a weather forecaster might use to predict the probability of rain. If he says there is a 90% probability of rain, he means that it's extremely likely that it will rain, although there's a slight chance that it won't. If he says there is a 60% probability of rain, he means that the chances of rainfall are not as great, although he feels that it is slightly more likely to rain than not. Similarly, if he gives a 40% probability he thinks that there probably won't be any rain, but there are four chances out of ten that there might be. If he gives a 10% probability, it is very unlikely that it will rain, but there is a slight chance. So, in the same way, when you see each pair on the recognition test I want you to give me a number between one hundred and zero. Give a large number when you think that it is probable that the pair has been changed, and a small number when you think that it hasn't been changed.

I'll always warn you when a recognition test is coming up. You'll have as much time as you want on the recognition test to give your response.

Do you have any questions?

Reminder Instructions

"This is a recognition test. Respons to each pair with a number between zero and one hundred. Remember, if you think that it's probable that the pair has been changed, give a large number. If you think that the probability is small that it's been changed, give a small number."